Attention Weights as a Window into How Crime Narratives Shape Subjective Assessments of Guilt

Anonymous ACL submission

Abstract

Neural network attention mechanisms have led to performance gains on a wide variety of tasks in natural language understanding, and they have also provided clues as to how these complex networks represent the language they process, opening up new avenues for linguistic investigation. In this spirit, we apply neural networks with attention to the task of modeling human judgments about guilt based on short crime narratives. The networks are successful at the task, and probing their attention weights helps illuminate the ways in which they make use of markers of certainty and uncertainty in different crime narratives, suggesting new hypotheses about how crimes might be more effectively reported in the news media.

1 Introduction

Deep learning models are increasingly applied to language tasks not only to develop technologies but also to derive new insights into language use. These more scientific applications raise the question of whether the models are processing data in a motivated way; their "black box" nature is often an obstacle to answering this question (Alishahi et al., 2019). Recently, attention mechanisms, which capture the strength of association between components of these networks (Bahdanau et al., 2015; Luong et al., 2015), have not only improved performance, but also increased interpretability, potentially opening the door to new linguistic applications of these models ([ek: paper that challenges this idea a little:](Serrano and Smith, 2019)).

In this paper, we apply neural networks with attention to a task that has both linguistic and societal import: predicting whether the reader of a short narrative crime report will conclude that the main subject is guilty or innocent. Our networks achieve strong performance on this task, as defined by the dataset of Kreiss et al. (2019), which leads us to explore their learned parameters by inspecting how they make predictions for new, carefully controlled examples. We find that the networks make robust use of markers of certainty and uncertainty, as well as discourse connectives that signal contrast. In addition, by systematically removing this language from test examples, we see that these markers have a large impact where the evidence is weak but have little or no impact where the evidence is strong. To the extent that humans adopt similar reading strategies, this can inform how the news is reported, in that it gives us insights into the contribution of words like *allegedly* and *possibly* in shaping readers' construal of newspaper articles (Erickson et al., 1978; Jensen, 2008).

2 Related Work

Our work draws on prior research concerning the relationship between language and assesments of guilt, as well as work seeking to use neural networks to inform linguistic theory.

2.1 Predicting Guilt

The challenge of predicting guilt judgments from text sources has not yet received a lot of attention. However, Fausey and Boroditsky show that using agentive language increases blame and financial liability judgments people make. Their results suggest that even subtle linguistic changes in crime reports can shape people's judgments of the events. More recent work has focused on predicting guilt verdicts from the Supreme Courts in the Philippines (Virtucio et al., 2018) and Thailand (Kowsrihawat et al., 2018) on the basis of facts presented and legal texts. Kowsrihawat et al. argue for a recurrent neural network with attention to make these predictions. This work is not concerned with the linguistic basis of subjective moral guilt judgments, but Court guilt verdicts on the basis of a given law text. It is our goal to train a network on

subjective guilt ratings and then probe its parameters with respect to their linguistic basis.

2.2 Interpretation of Hedges

In the literature, researchers have categorized and labeled (un)certainty markers in numerous ways, e.g., (Lakoff, 1972; Prince et al., 1982; Brown et al., 1987). For a summary, we refer readers to Fraser 2010. In this work, we use "hedge" broadly, as an umbrella term for all subclasses of uncertainty markers, including any marker that introduces uncertainty "within the propositional content" (e.g., "His feet were **sort of** blue" (Prince et al., 1982)) or "in the relationship between the propositional content and the speaker" (e.g., "**I think** his feet were blue" (Prince et al., 1982)).

There is extensive prior literature on how hedges affect perceptions of the speaker and proposition (Erickson et al., 1978; Durik et al., 2008; Bonnefon and Villejoubert, 2006; Rubin, 2007; Jensen, 2008; Ferson et al., 2015). The resulting picture is highly varied. These studies suggest that hedges affect people's judgments of credibility in differing ways. For example, an increase in the number of hedges decreases the credibility of witness reports (Erickson et al., 1978), but, at the same time, increases the trustworthiness of journalists and scientists (Jensen, 2008). Additionally, the interpretation of hedges is very context dependent (Bonnefon and Villejoubert, 2006; Durik et al., 2008; Ferson et al., 2015) and shows high individual variation (Rubin, 2007; Ferson et al., 2015). We investigate how neural network predictions about human judgments change when we manipulate the presence of those hedges in crime stories.

In addition, computational and theoretical work on veridicality has revealed that not all hedges serve merely to reduce speaker commitment. For instance, attitude predications like "The company reported S" and "They said S" are often used to convey evidence sources. In such utterances, the speaker might wish to appear fully committed to the embedded content "S" (Simons, 2007; de Marneffe et al., 2012; White and Rawlins, 2018; White et al., 2018). von Fintel and Gillies (2010) show that similarly evidential readings arise for epistemic "must". These findings show how complex these markers are pragmatically and highlight the value of usage-based studies of them.

2.3 Gaining Linguistic Insights Through Neural Networks

Neural networks were originally motivated largely by questions in cognitive science (Pater, 2019), especially concerning issues of representation and learning (e.g., (Rumelhart and McClelland, 1986; Tesar and Smolensky, 2000)). Today, they are more closely associated with engineering efforts, but there is increasing interest in returning to these models' cognitive motivations and in using them to try to gain insights about linguistic phenomena. The incredible representational power of these networks has long been an obstacle to such work, but recent efforts are overcoming these obstacles. For instance, Ribeiro et al. (2016) and Koh and Liang (2017) develop methods for probing complex models to understand which features are guiding their predictions. Attention weights are increasingly playing a role in such efforts. For instance, Jawahar et al. (2019), Clark et al. (2019) and Tenney et al. (2019) present evidence that BERT language models systematically encode core aspects of linguistic structure. We present a similar argument here, but focused on usage-based patterns. Finally, Linzen et al. (2016) explicitly argue for applying methods from psycholinguistics to trained neural networks, treating them as participants whose performance we can observe and learn from (see also Gulordava et al. 2018; Futrell et al. 2019).

3 Dataset

For training and testing the model, we use the Annotated Iterated Narration Corpus (AINC) collected by Kreiss et al. (2019).[1] The corpus was created for investigations on how news stories change when they are propagated from one person to another.

First, the authors created five stories of approximately 850 words, which are called the *seed* stories. All of them report a crime and an arrest of one or more suspects. Each of these five seed stories exists in a weak and a strong evidence condition. The stories are identical up to the last phrase, which then either raises doubts about the arrest or emphasizes its validity. For example, if a suspect was arrested on the basis of camera footage, this footage was described as being of "very poor quality" in the weak evidence condition and "very high

[1] https://github.com/elisakreiss/iterated_narration

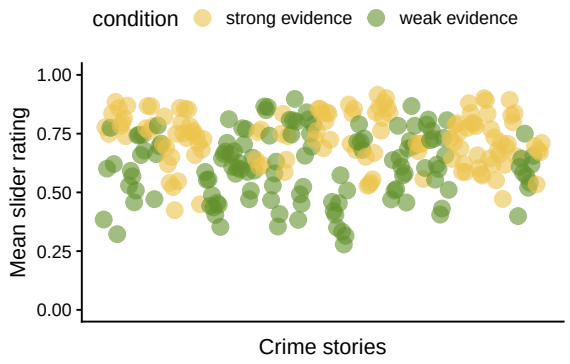

Figure 1: A point represents the mean subject guilt rating for a story in the corpus. They are color-coded with respect to their condition.

quality" in the strong evidence condition. Table 1 provides a complete example.

Each story was given to a participant who was asked to read and afterwards reproduce it. The reproduced story was then given to the next participant who again reproduced it. This pattern was repeated for 5 generations of participants. The data collection therefore followed the transmission-chain paradigm, introduced by (Bartlett, 1932). Following this schema, each story and condition was reproduced in 5 chains over 5 generations, resulting in 250 reproductions, and therefore 260 stories overall.

After the corpus collection, Kreiss et al. annotated the corpus with human judgments. The questions were primarily related to different aspects of guilt perception but also, for example, perceived subjectivity of the story writing. Participants were recruited on Amazon Mechanical Turk and indicated their response on a continuous slider, here underlyingly coded as ranging from 0 to 1. Most interestingly for this project, they asked "How likely is it that the suspect is / the suspects in the crime are guilty?" Each story received approximately 20 ratings, ranging from 0 to 1. For the purposes of this work, we consider the mean rating for each story as its guilt judgment label. The labels of each story are shown in Figure 1. In contrast to the raw ratings, the means only range from 0.27 to 0.92.

In summary, the AINC is a corpus of reproduced news stories, annotated by human guilt judgments. Since the stories originated from only 5 unique stories (each in 2 conditions), a lot of information is shared between single data points. Despite this similarity, the range of guilt judgments is still high, alluding to subtle differences that trigger this variance. We now turn to the question of whether a deep learning model can learn to predict the assessments of guilt the corpus provides.

4 Model

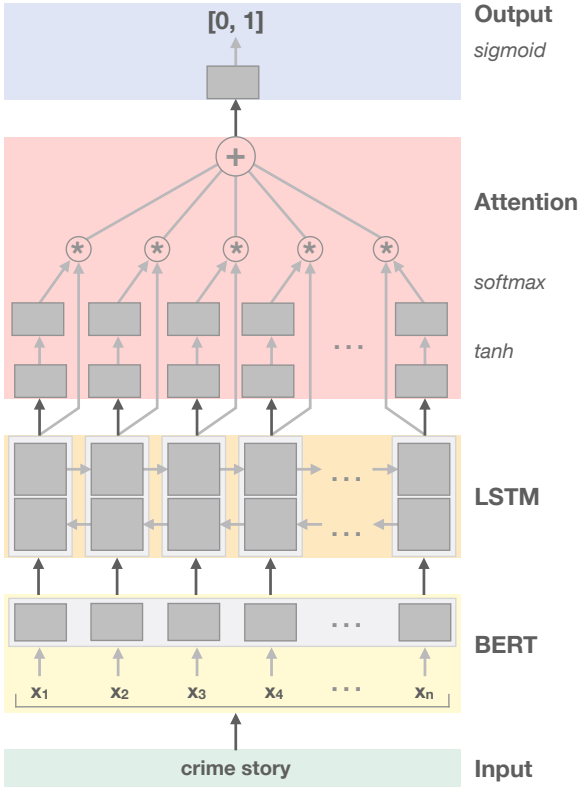

Figure 2: Model architecture. The BERT layer produces a sequence of context-dependent representations. These are the inputs to the bidirectional LSTM, which fine-tunes these representations to our task. The LSTM outputs feed into the attention layer, which produces a summary representation of the full text that is weighted by each word's contribution to the final predictions.

Our model is based on the one proposed by Lin et al., which is fundamentally a bidirectional LSTM with attention mechanisms applied to its outputs. We chose this model primarily because its attention mechanisms seem especially promising for introspection, as they essentially provide a weight for each word in the input, and this weight controls how much each word contributes to the network's final predictions. The overall architecture is summarized in Figure 2.

The only major adjustment we made to the model is that we replaced the GloVe word embeddings (Pennington et al., 2014) by BERT representations (Devlin et al., 2019). Whereas GloVe pro-

On August 30 2018, 30,000 internet users in Spain received an email with the subject line 'Immediate response required - copyright infringement'. In these emails, the recipients were accused of illegally distributing pornographic content and were instructed to pay $11,500 by the end of September to avoid a lawsuit. Searching for the email address online leads to an official law firm website located in Barcelona. Shortly after the emails were sent out, the law firm disowned them, calling them a hoax. So far, police officials estimate that approximately 5% of the recipients have paid the money. This sums to more than $17 million that have been paid to the scammers. Now, nearly 4 weeks later, the police have tracked down the internet cafe from which the emails were sent. After observing the cafe's camera footage, the police arrested a man and a woman (both 52) who were allegedly responsible for the scam. If found guilty, they face up to 9 years in prison.	
(*strong evidence condition*) Moreover, news outlets recently acquired access to the camera footage and found that the video material is of very high quality and therefore undoubtedly reliable.	(*weak evidence condition*) However, news outlets recently acquired access to the camera footage and found that the video material is of very poor quality and therefore potentially unreliable.

Table 1: Example of a seed story in the two conditions from the Annotated Iterated Narration Corpus (AINC).

vides a single embedding for each word, with no sensitivity to the context in which it occurs, BERT representations vary by syntactic context. Devlin et al. report that using these embeddings boosts performance in all of the 11 natural language tasks they consider, and we saw comparable improvements when we switched from GloVe to BERT. To access pretrained BERT parameters, we used the Hugging Face toolkit.[2]

We allow BERT to tokenize the input string according to its internal tokenization method (Wu et al., 2016), to make maximal use of its own pretrained embedding. These token representations are processed by BERT using its pretrained Transformer parameters (Vaswani et al., 2017), yielding a sequence of contextual representations for them. These are fed into a bidirectional LSTM layer in which each's cell's output has dimension 200. The two representations at each step are concatenated to form the LSTM layer output for each token.

For the attention layer, we follow the design of Lin et al.: the LSTM outputs are organized into a matrix H of dimension $n \times 400$, where n is the token length of the current sequence, and we apply a dense layer with parameters W (dimension 50×400) and a $\tanh$ activation to create a matrix A of 50-dimensional representations for each token:

$$A = \tanh(WH^\top) \tag{1}$$

The matrix A is further compressed to the attention weight vector $\mathbf{a_w}$ of size 1×50, with a softmax applied so that the weights sum to 1:

$$\mathbf{a_w} = \mathrm{softmax}(WA) \tag{2}$$

Lin et al. actually generalize this attention operation to return a matrix with r attention weights for each token. We set $r = 1$ to keep the number of parameters low and increase the interpretability of the network.

Finally, we take the dot product of the LSTM hidden state matrix H and the just obtained vector a_w to obtain the attention output vector $\mathbf{a}_{\text{out}}$ of size 1×400:

$$\mathbf{a}_{\text{out}} = \mathbf{a_w}H \tag{3}$$

This is the output of the attention module. Intuitively, the attention weights capture the relative importance of each token with regard to the final prediction, and $\mathbf{a}_{\text{out}}$ synthesizes all of these weighted contributions into a single vector.

Finally, the logistic regression module makes the prediction of guilt. For this step, $\mathbf{a}_{\text{out}}$ is linearly transformed into a single number $p = \mathrm{sigmoid}(\mathbf{a}_{\text{out}})$. The sigmoid function ensures that the prediction lies between 0 and 1, just like the restriction on the guilt judgments it is trained on.

Overall, the model has $111,054,742$ trainable parameters. This sounds mismatched with our very small dataset, but only 0.02% of these parameters are in the attention and output modules and 1.40% are in the LSTM module. The rest of the parameters (98.59%) are the pretrained weights in the BERT word embedding module.

[2] https://github.com/huggingface/pytorch-transformers

In summary, the model receives a crime story as an input. BERT word embeddings for the story feed into a bidirectional LSTM. The attention module computes the weighted sum over the LSTM's hidden states. The resulting attention vector is fed through a linear layer and a sigmoid function to ensure that the prediction lies between 0 and 1.

5 Experiment

Our initial goal is to assess the extent to which our bidirectional LSTM with self-attention (as described in Section 4) can predict human guilt judgments from news stories. Assuming the model succeeds, we can then probe its internal representations for linguistic insights.

5.1 Optimization

To begin with, we held out 26 randomly selected stories (from the 260 stories in total) as the final test set. The remaining 234 stories were then used for model training and validation, which was done using 10-fold cross validation. The cross-validation results inform us about the model variation observed across various training/validation splits; it is quite likely that, given the small size of the dataset, this variation could be quite high.

In each step of the 10-fold cross-validation, the model was trained on 206/207 stories and 23/24 were held out for dev-testing. As noted above, this guilt rating is the mean participant guilt rating obtained from the previously described data collection and annotation.

For training, the model uses mean squared error (MSE) as the loss function, stochastic gradient descent as the optimizer, and a learning rate of 0.1. The model was implemented using PyTorch (Paszke et al., 2017). It was trained for 30 epochs, for each of the 10 training/dev-test configurations in the cross-validation.

5.2 Results

Figure 3 shows the MSE loss for each training epoch. We show the dev-set loss for our model (orange) as well as the train-set loss (blue). In addition, to provide context for the results, we include the loss for a dummy regressor that simply predict the mean of the training data labels for all cases.

The training loss approaches 0 toward the end of the training in all cross-validation configurations, indicating model convergence. Crucially, our model's dev-set performance is always substantially better than the baseline model, indicating that the model is indeed learning from the dataset. That said, there is a high amount of variation between the different cross-validation steps, with the baseline actually proving competitive in some folds. This seems an inevitable consequence of our small dataset, but the model clearly has gotten traction on the problem overall.

The MSE loss alone is not sufficient to assess how well the model actually learns to predict the underlying distribution. Figure 4 shows the correlation between the actual target labels (on the x axis) and the model predictions (on the y axis) for one of the cross-validation folds. Before training (left), the models are undifferentiated. After training (right), the model predictions (blue) are highly correlated with the true labels ($r = 0.85$), and the MSE is small (0.007).

Qualitatively, these plots appear very similar throughout all cross-validation configurations and can be compared in Figure 8 and 9 in the Appendix. Quantitatively the correlation on the dev-test set does show variation, driven mainly by outliers.[3] However, overall, the mean correlation between the model prediction and the human judgment on the testing set across all cross-validation steps is 0.68. When we collapse over all cross-validation folds and examine the loss after training, the difference in loss between the model predictions and baseline is significant ($p < 0.0001$).[cp: Which test was used to get this p-value?]

As a final performance evaluation, we examine the target–prediction correlation on the held-out test set (Figure 5). For predictions on the test set, we used the final model weights obtained by the first cross-validation configuration. This setup was chosen because of the low dev-set loss and the high correlation of the dev-set with the target labels. This was the only evaluation that was performed on this held-out test set. The Pearson correlation on the held-out test set is still high (0.84) and almost identical with performance on the dev-set for this fold. This high correlation, and the fact

[3] [ek: Note on professor seed as outlier: Story about sexual harrassment allegations against male professor; only story that has clear gender split between victim(s) and suspect; Erickson 1978 show that powerless style (which includes hedges) affects credibility ratings dependent on whether witness is same or other sex]

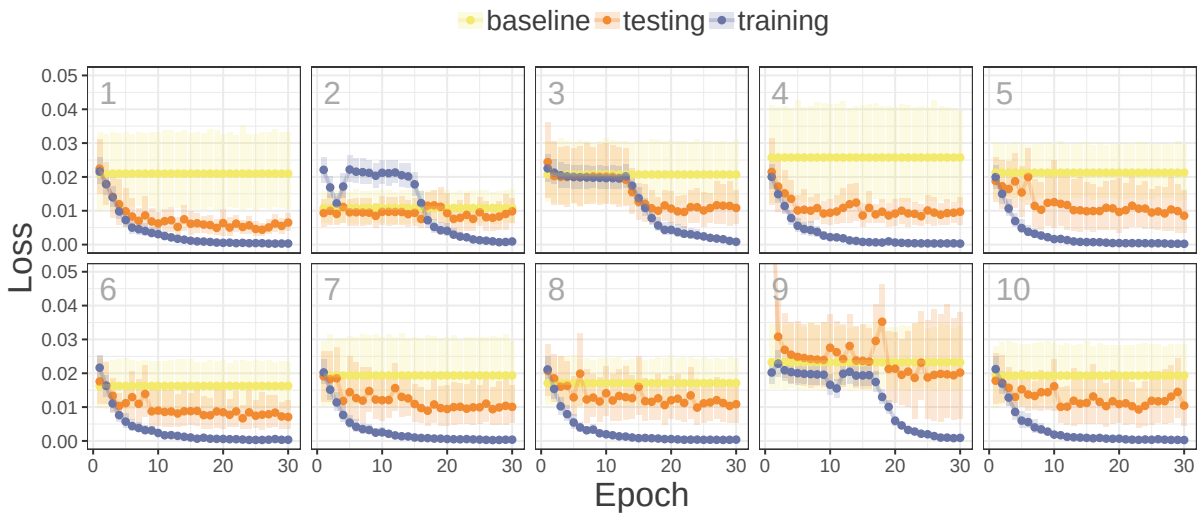

Figure 3: Loss (mean squared error) over epochs (x axis), faceted over cross-validation configurations. The performance of the model on the training set (in blue) approaches zero. The performance of the model on the dev-test set (in orange) generally outperforms the baseline (in yellow).

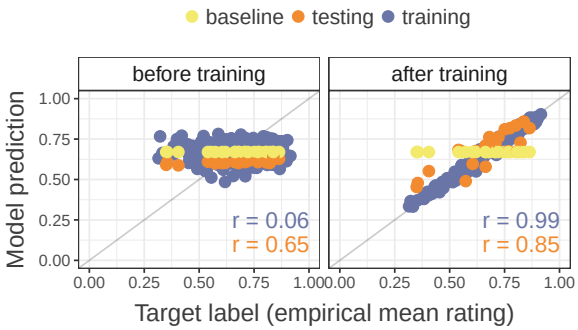

Figure 4: Correlation between target label (x axis) vs. model prediction (y axis) before and after training.

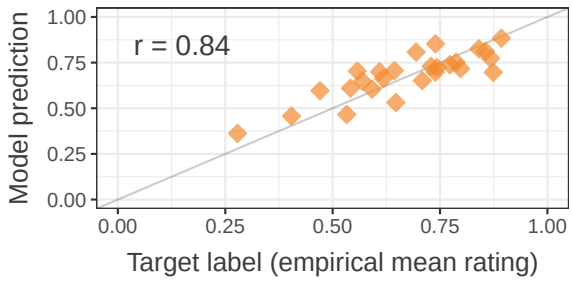

Figure 5: Testing label (x axis) vs. model prediction (y axis) after training on the held out test set (26 data points) using the parameter settings obtained after the 30th epoch from the first cross-validation. Pearson correlation of 84%.

that the high correlation reproduces with the held-out testing data, indicates that the model was able to learn to generalize.

6 Model analysis

We have established that the proposed model can predict human guilt judgments when given a crime story. This result invites us to ask the question of whether there are patterns underlying these predictions that can provide higher-level insights into how langauge is construed in these criminal contexts.

6.1 Visualization

We begin by focusing attention on the learned attention weight vector $\mathbf{a_w}$ in equation (2). Since the softmax forces the sum of the weights to be 1, we cannot interpret the weights on their own for each word or across stories. Instead, the relevance lies in the differences between words and phrases within each story, and patterns of similarities between stories.

To investigate what might affect model predictions, we ran the model again on the final test data. Figure 6 displays three of these stories with their attention weight distribution. We see that the model seems to focus on phrases that explicitly describe uncertainty about the evidence (see Figure 6C). If present, these cues usually outweigh the rest of the story. This suggests that, if there is an explicit claim that affects the evidence of the suspect's guilt, it is considered as the most important source to inform guilt judgment.

Additionally, peaks occur on words and phrases which communicate turning points in a story, such

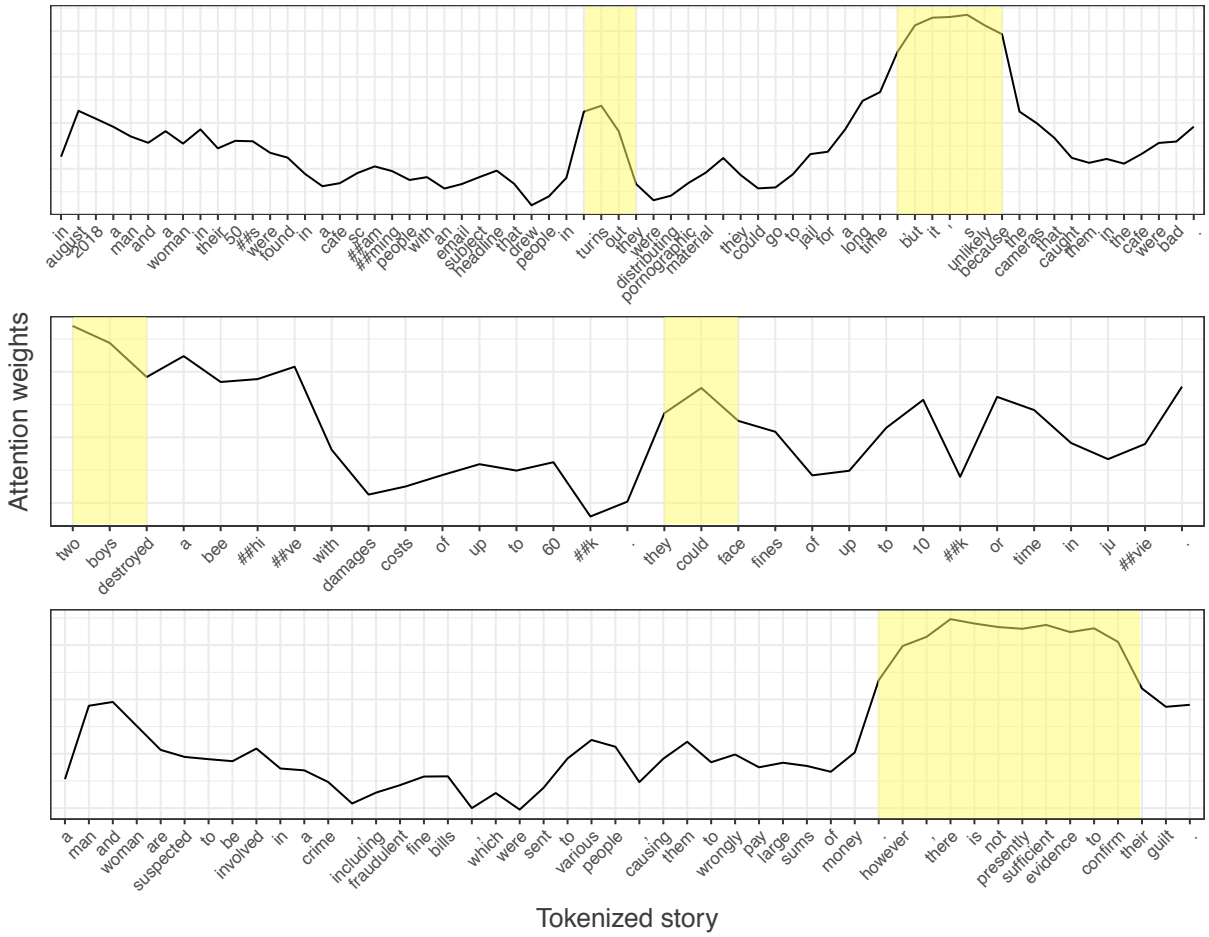

Figure 6: Visualization of the attention weights (y axis) for a tokenized story (x axis) from the test set. Because we took the softmax over the attention weights, the scale of the y axis is irrelevant. Areas of high attention weight are marked in yellow and correlate with markers of (un)certainty as well as markers of contrast. [cp: The panel needs A, B, and C marks. I thought about adding them to the PDF image, but there isn't really room in it.]

as "however", "but", "even though", and "it turns out". This can be seen in all three stories in Figure 6. Those phrases could be relevant because they usually indicate a contrast to a prior narrative. Since those markers mostly follow reports of arrests, they might be strongly correlated with objections to those arrests which would in turn influence guilt perceptions.

Figure 6B shows a case where the model seems to find a simple declarative ("two boys destroyed") to be relevant for the final prediction. This is especially interesting, since declaratives on their own do not generally communicate guilt-related information. However, they are very important for guilt judgments because they do not allow any uncertainty about the association between crime and suspect.

In summary, the visualization of the attention weights contribute further evidence that the model learns meaningful patterns in the data.

6.2 Qualitative Analysis

Figure 6 provides evidence for the claim that the model picks up on meaningful patterns in the corpus. But does the model make reasonable predictions on newly constructed data points? The original corpus started out with five different crime stories. Each of these stories occurred in two conditions – one suggesting that the evidence that led to the arrest was weak, and the other that the evidence provided a strong case. Additionally, the stories were filled with uncertainty markers such as "allegedly" or "(un)likely", which we expect to have an impact on guilt judgments. Howver, the corpus cannot inform us about this relationship.

We can, though, use the model to predict a guilt judgment for each of the original stories without these uncertainty markers/hedges. Thus, we rewrote those 10 original stories into versions without any uncertainty markers. They remained

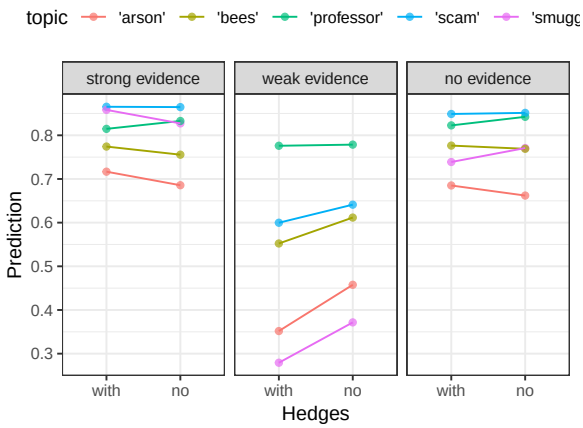

Figure 7: Model predictions on the relationship between uncertainty markers and evidence statements. [ek: fix caption and make figure more readable]

as close to the original as possible, while still remaining grammatical. Figure 7 summarizes this analysis. The results suggest that uncertainty markers have different effects on guilt prediction in the two evidence conditions. When the evidence is strong, removing all uncertainty markers does not affect the guilt judgments. However when the evidence is weak, removing those hedges increases the guilt judgment. This has an intuitive interpretation: when evidence already overwhelmingly speaks for the suspect's guilt, it outweighs the hedges; when the evidence is questionable, other sources of uncertainty are considered to inform a final judgment.

Notably, if the evidence manipulation is excluded, the model predicts guilt judgments in the range of the strong evidence condition. In other words, the model assumes that additional justification for an arrest does not change the model's prediction of a guilty judgment. However, if reasons to question the arrest are given explicitly, this does affect the model's prediction. When we remove the hedges again, we cannot see a common structure to the change in ratings (possibly more similar to the strong evidence ratings though).[cp: I am not sure I understand this last sentence. Could it be expanded?]

7 Discussion

[ek: past tense!] The way a crime and arrest is presented in a news article affects how readers perceive the suspect's guilt. In this paper we showed that a recurrent neural network with self-attention can predict readers' guilt judgments. By visualizing the attention weights, we find that explicit evidence descriptions and phrases suggesting a turn of events influence predictions.

To gain linguistic insights into the neural network, we investigated its change in predictions when we exclude the hedges and information about the evidence from the stories. First of all the model makes similar predictions in the case where the evidence leading to an arrest is strong/convincing and the case where the evidence is not specified. [ek: One possible interpretation of this result is that the model learns that there is an implicature of strong evidence if the evidence is underspecified?] The model predictions are affected the most when the weakness of the evidence is mentioned explicitly.

Furthermore, we found that an exclusion of hedges in the stories only seems to affect predictions of stories in the weak evidence condition, where suspects are rated more guilty when there are no hedges. However, if the evidence is strong or underspecified, the model predictions do not seem to be affected. Possibly, in case of strongly perceived evidence, the model has learned that the hedges become irrelevant.

These results point to promising avenues for future investigations of hedging, particularly in the field of guilt perception.

[ek: how we perceive stories affects how we reproduce them; relevance for iterated chains]

[ek: Conclusion: propose a network that can predict guilt judgments and might be used to inform hypotheses for insight on linguistic features that determine guilt perception...]

References

Afra Alishahi, Grzegorz Chrupała, and Tal Linzen. 2019. Analyzing and interpreting neural networks for NLP: A report on the first BlackboxNLP workshop. *Journal of Natural Language Engineering*, 25(4):543–557.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations 2015*.

Frederic C Bartlett. 1932. Remembering: An experimental and social study. *Cambridge: Cambridge University*.

Jean-François Bonnefon and Gaëlle Villejoubert. 2006. Tactful or doubtful? expectations of politeness ex-

plain the severity bias in the interpretation of probability phrases. *Psychological Science*, 17(9):747–751.

Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does bert look at? an analysis of bert's attention. In *BlackBoxNLP@ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Association of Computational Linguistics*, Stroudsburg, PA. Association for Computational Linguistics.

Amanda M Durik, M Anne Britt, Rebecca Reynolds, and Jennifer Storey. 2008. The effects of hedges in persuasive arguments: A nuanced analysis of language. *Journal of Language and Social Psychology*, 27(3):217–234.

Bonnie Erickson, E Allan Lind, Bruce C Johnson, and William M O'Barr. 1978. Speech style and impression formation in a court setting: The effects of "powerful" and "powerless" speech. *Journal of Experimental Social Psychology*, 14(3):266–279.

Caitlin M Fausey and Lera Boroditsky. 2010. Subtle linguistic cues influence perceived blame and financial liability. *Psychonomic Bulletin & Review*, 17(5):644–650.

Scott Ferson, Jason O'Rawe, Andrei Antonenko, Jack Siegrist, James Mickley, Christian C Luhmann, Kari Sentz, and Adam M Finkel. 2015. Natural language of uncertainty: numeric hedge words. *International Journal of Approximate Reasoning*, 57:19–39.

Kai von Fintel and Anthony S. Gillies. 2010. Must ... stay ... strong. *Natural Language Semantics*, 18(4):351–383.

Bruce Fraser. 2010. Pragmatic competence: The case of hedging. *New approaches to hedging*, 1(1):15–34.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Jakob D Jensen. 2008. Scientific uncertainty in news coverage of cancer research: Effects of hedging on scientists' and journalists' credibility. *Human communication research*, 34(3):347–369.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894, International Convention Centre, Sydney, Australia. PMLR.

Kankawin Kowsrihawat, Peerapon Vateekul, and Prachya Boonkwan. 2018. Predicting judicial decisions of criminal cases from thai supreme court using bi-directional gru with attention mechanism. In *2018 5th Asian Conference on Defense Technology (ACDT)*, pages 50–55. IEEE.

Elisa Kreiss, Michael Franke, and Judith Degen. 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.

George Lakoff. 1972. Hedges: A study in meaning criteria and the logic of fuzzy concepts. In *Papers from the Eighth Regional Meeting of the Chicago Linguistic Society*, pages 183–228. Chicago Linguistic Society. Reprinted in Journal of Philosophical Logic, 1973, 2: 4, 458–508, and in D. Hockney et al. (eds.). Contemporary research in philosophical logic and linguistic semantics. Dodrecht: Fortis, 221–271.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Joe Pater. 2019. Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*.

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

Ellen F Prince, Joel Frader, Charles Bosk, et al. 1982. On hedging in physician-physician discourse. *Linguistics and the Professions*, 8(1):83–97.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101. Association for Computational Linguistics.

Victoria L. Rubin. 2007. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 141–144, Rochester, New York. Association for Computational Linguistics.

D. E. Rumelhart and J. L. McClelland. 1986. On learning the past tenses of English verbs. In David E. Rumelhart, James L. McClelland, and CORPORATE PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2*, pages 216–271. MIT Press, Cambridge, MA, USA.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.

Mandy Simons. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua*, 117(6):1034–1056.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Bruce Tesar and Paul Smolensky. 2000. *Learnability in optimality theory*. Mit Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Michael Benedict L Virtucio, Jeffrey A Aborot, John Kevin C Abonita, Roxanne S Aviñante, Rother Jay B Copino, Michelle P Neverida, Vanesa O Osiana, Elmer C Peramo, Joanna G Syjuco, and Glenn Brian A Tan. 2018. Predicting decisions of the philippine supreme court using natural language processing and machine learning. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 130–135. IEEE.

Aaron Steven White and Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*, Amherst, MA. GLSA Publications.

Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2018. Lexicosyntactic inference in neural models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4717–4724, Brussels, Belgium. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. ArXiv:1609.08144.

8 Appendices

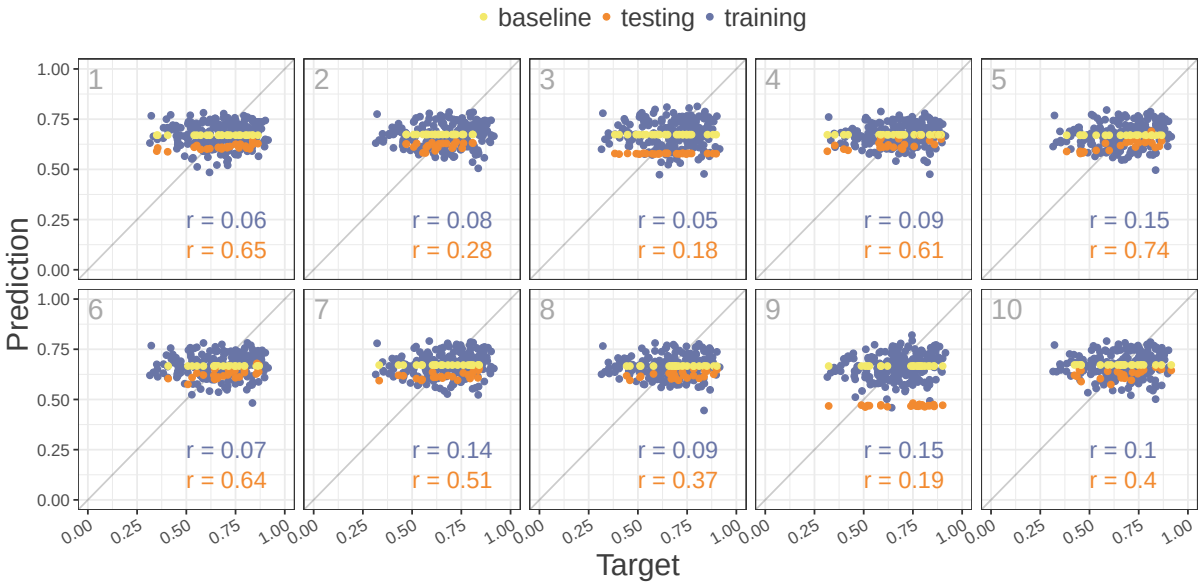

Figure 8: Testing label (x axis) vs. model prediction (y axis) before training; faceted over cross-validation configurations.

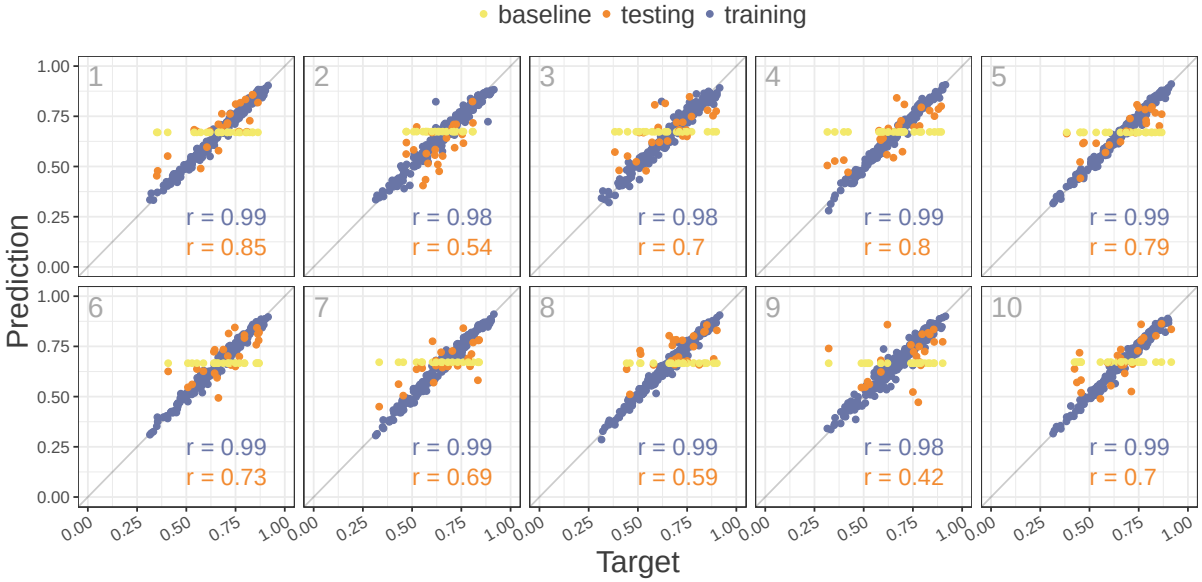

Figure 9: Testing label (x axis) vs. model prediction (y axis) after training; faceted over cross-validation configurations.