Veridicality, Negation-Raising, and Fine-Grained Inference

Will Gantt, Ruoyang Hu, and Ben Kane

Abstract

We investigate the relationship between the belief and desire relations expressed by certain verbs and those verbs' tendency to license negation-raising and veridicality inferences. To do this, we collect a preliminary pilot dataset of human judgements on a natural language inference (NLI) task designed to target the belief and desire properties of different English verbs. We then train an LSTM-based neural NLI model with RoBERTa subword embeddings on existing datasets of negation-raising and veridicality inference judgments and evaluate its ability to correctly make the inferences in the new task. Our initial findings, based only on the small set of judgments from this pilot (fewer than 100 examples), are inconclusive.

1 Introduction

Clause-embedding predicates that trigger inferences about the factuality of propositions in the embedded clause are said to be *veridical*. For example, both (1a) and (1b) trigger the inference (2a).

(1) a. Jo liked that Bo left.
    b. Jo proved that Bo left.

(2) a. Bo left.
    b. Bo didn't leave.

The verbs *like* and *prove* are thus veridical. However, when the predicate in the main clause is negated, the inference might change.

(3) a. Jo didn't like that Bo left.
    b. Jo didn't prove that Bo left.

The negated *like* in (3a) can still trigger inference (2a), whereas in (3b), we cannot tell whether the embedded clause ("Bo left") is true or not, and so cannot infer either (2a) or (2b). Verbs such as *like* for which one can still infer the truth of its complement under negation are examples of *factives*—predicates that do not merely assert the truth of their embedded clause, both presuppose it.

There is already a considerable literature on veridicality and factivity (White and Rawlins, 2018; Anand and Hacquard, 2014; White, 2020). Another class of inference, negation-raising (or simply *neg*-raising), has also recently been drawing more research interest (Gajewski, 2005; Gajewski, 2007; Cohen, 2017; An and White, 2019). Neg-raising occurs when negation on the predicate of a main clause can be interpreted as negating its embedded clause. For example, Neg-raising is triggered in (4a) but not (4b).

(4) a. Jo didn't think that Bo left $\rightsquigarrow$ Jo thought that Bo didn't leave.
    b. Jo didn't proved that Bo left $\not\rightsquigarrow$ Jo proved that Bo didn't leave.

Previous studies that examine veridicality and neg-raising inferences have considered them as lexical properties that may help explain certain syntactic phenomena (Theiler et al., 2019; Uegaki et al., 2017; White, 2019b), such as the types of clauses that predicates select for. However, there is limited literature showing how, if at all, these properties interact, which is one of the objectives of our work.

As demonstrated above, the verb *prove* is veridical but not neg-raising, which is the same for *like* and *know*. It appears especially difficult to find a verb that is both factive and neg-raising, since by definition a verb is factive iff it is veridical and NP V S presupposes S—that is, if both NP V S and NP NOT V S entail S. This would seem to explicitly contradict the definition of neg-raising, which requires that NP NOT V S entails NP V NOT S.

A possible explanation for this is that veridical and factive predicates like *like* and *know* describe a certain relation between a person (e.g. "Jo") and a proposition ("Bo left"). Considering *like* (a factive), for instance, it seems to communicate two things. First, it plainly expresses a relation of *desire*: the *liking* articulates Jo's satisfaction at Bo's having left. Second, it expresses a *belief* relation: the liking also presupposes Jo's belief that Bo left. When *like* is negated (1a), the desire relation is negated (or "backgrounded"), though the belief relation is sustained ("foregrounded"). By contrast, *know* seems to have a weaker desire component but a stronger belief component.

Our hypothesis is that these belief and desire factors may help explain both the veridicality and the neg-raising tendencies of verbs. To test this, we design a natural language inference (NLI) task that uses negation to determine whether the belief and desire components are foregrounded or backgrounded for different predicates and deploy it on Amazon Mechanical Turk. In these experiments, we consider pairs of sentences—an antecedent and a consequent. The antecedent sentence includes a clause-embedding predicate and its complement. The consequent sentence asks one of two questions aimed at determining the strength of the belief component or the desire component. We also include negated variants of the antecedent and consequent sentences, in order to identify how these components may interact with negation. Lastly, we implement a set of neural NLI models trained on veridicality and neg-raising data and evaluate their performance on our new task.

The rest of the paper is outlined as follows: §2 describes our data collection and cleaning procedures and our model architecture; §3 presents the experiments conducted with the model and the results; §4 reviews some related work; and §5 outlines our intentions for future work, describes our individual contributions to this report, and concludes with some acknowledgments.

2 Methodology

2.1 Data Collection

The first major part of this project was the collection of two small pilot datasets intended to test foregrounded/backgrounded belief and desire components of a small set of verbs chosen from the literature. The first of these pilots concerns emotive and cognitive verbs (e.g. "love" and "know" respectively), while the second pilot concerns communicative verbs (e.g. "say" and "tell"). These pilots are intended to help us test the validity of this new experimental setup before conducting a bulk study using the full set of 800+ verbs from the MegaAcceptability dataset [1].

To obtain this data, we carry out Amazon Mechanical Turk (MTurk) data collection

[1] http://megaattitude.io/projects/mega-acceptability/

using the "semantic bleaching" method described in (White, 2019a). Specifically, the syntactic context of each verb of interest is manipulated while fixing other constituents to low-content placeholders (e.g. "someone did something"). At the time of writing, we've collected the data for the first pilot, and have prepared the code for generating the second pilot MTurk task, although have not yet collected the data.

To create the MTurk experiments, we use a fork of the Ibex library created by Aaron White[2]. This requires a CSV file containing the instruction set, list of items, and any other content to be included in the experiment. We used a slightly modified version of the instruction set from previous NLI studies by Aaron White (since the instructions are generic), along with a few practice questions, and "sanity check" questions used to ensure quality of responses (which were interspersed with the actual items).

For our first pilot, we chose 12 predicates from the literature, with 6 of these being cognitive verbs and 6 being emotive verbs. To keep our pilot size small and reduce the amount of possible pragmatic factors which might confound our analysis of the task itself, we look only at predicates in past tense, 3rd person, and in a "that S" frame. For example, given the predicate "doubt", we would obtain the basic antecedent sentence "A particular person doubted that a particular thing happened". Each antecedent is paired with a consequent of the form "They believed that that thing happened" and "They wanted that thing to have happened". Lastly, we create all 4 possible negation configurations (with the negation taking on high scope in each case) between the antecedent and consequent sentences. While choosing the 12 predicates to use, we took care to ensure that the expected inference judgments based on the literature were approximately balanced for each consequent across the different negation configurations.

Based on the manipulations described above, we end up with 96 individual items. We distribute these items evenly across 4 MTurk "human intelligence tasks" (HITs), so that each HIT has 24 items. We ensure that each HIT has one instance of every combination of predicate and consequent, but shuffle the negation configurations for each item across the 4 HITs so as to avoid introducing a source of bias. Additionally, we add four sanity check questions to each HIT, where the consequent is either identical to the antecedent or the exact negation of the antecedent. Each HIT accepted 10 unique participants, and no participant was allowed to accept multiple HITs. Consequently, we ended up with 1,120 data points in the first pilot, including the sanity check questions.

Since our experiment is still in its preliminary stages and we are primarily interested in developing the inference architecture for the bulk study, we have not yet fully completed cleaning our data or recollecting the rejected data. Two forms of data cleaning will ultimately need to be performed: excluding any data from participants who answered the sanity check questions incorrectly, and excluding data with low inter-annotator agreement. A cursory review of the data indicated that the issue of failed sanity check questions results in data from a few participants being invalid, though the majority is unaffected.

The second pilot will be generated in a similar way using communicative predicates. One difficulty with creating the second pilot, which motivated us to separate it from the first pilot, is that the distribution of frames and inferences correlated with those frames seems to differ between communicative predicates and

[2]https://github.com/aaronstevenwhite/ibex

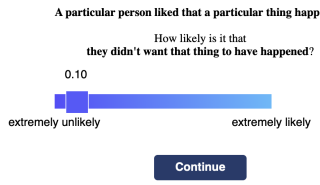

Figure 1: An example of an item from the first pilot with a negated "want" consequent.

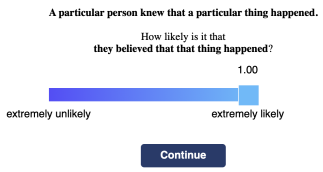

Figure 2: An example of an item from the first pilot with a positive "believe" consequent.

the cognitive/emotive predicates. In particular, the frames in "tell Y that X" vs. "tell Y to X" seem to be highly correlated with whether we get the "believe that X" or the "want X to happen" inference, and it's not clear whether we should limit the frames in the antecedents to just "that S" (allowing only the former antecedent) like the first pilot, or allow both. We're still considering the best way to address this difficulty before launching the second pilot.

Two example items from a HIT from the first pilot are shown in Figure 1 and 2.

2.2 Model

Our model takes two sentences as input, an antecedent and a consequent, and outputs a score corresponding to the likelihood that the latter is entailed by the former. We first obtain 768-dimension subword embeddings for each sentence from the RoBERTa base model (Liu et al., 2019). The subword embeddings for the antecedent and for the consequent are then fed separately as inputs to a single, one-layer LSTM. The final hidden states of the LSTM for the two sentences are concatenated and the resulting vector is used as input to a multi-layer perceptron, consisting of two fully-connected layers with ReLU activations and a linear output layer. We use mean squared error for our loss function.

Ordinarily, loss could be computed using the raw output from the MLP. However, the precise normalization procedures used to generate the MegaVeridicality and MegaNegRaising datasets differ, and result in different scales for the inference likelihood scores. Although MegaNegRaising and our pilot data use scores normalized to the range $(0, 1)$, MegaVeridicality does not. Thus, to allow joint training on these differently normalized data, we fit separate isotonic regressions for each dataset between the MLP output and the true scores for that dataset. The isotonic regression fits a monotonically increasing curve that effectively calibrates the MLP output to the scale of the dataset to which it is fitted. We describe the procedure for fitting the curve in more detail in §2.3.

The model was implemented in Python using PyTorch version 1.4, and we used the Scikit-Learn library's implementation of the isotonic regression model.[3]

2.3 Normalization

Since annotators may interact with the continuous scales from the MTurk task in different ways, the scores obtained in the pilot cannot be used directly by the model. First, we normalize the scores between the annotators for each HIT, and obtain a single score for each individual item. This is done using a mixed effects robust regression similar to that in (An and White, 2019), however we do not jointly

[3] https://scikit-learn.org/stable/modules/generated/sklearn.isotonic.IsotonicRegression.html

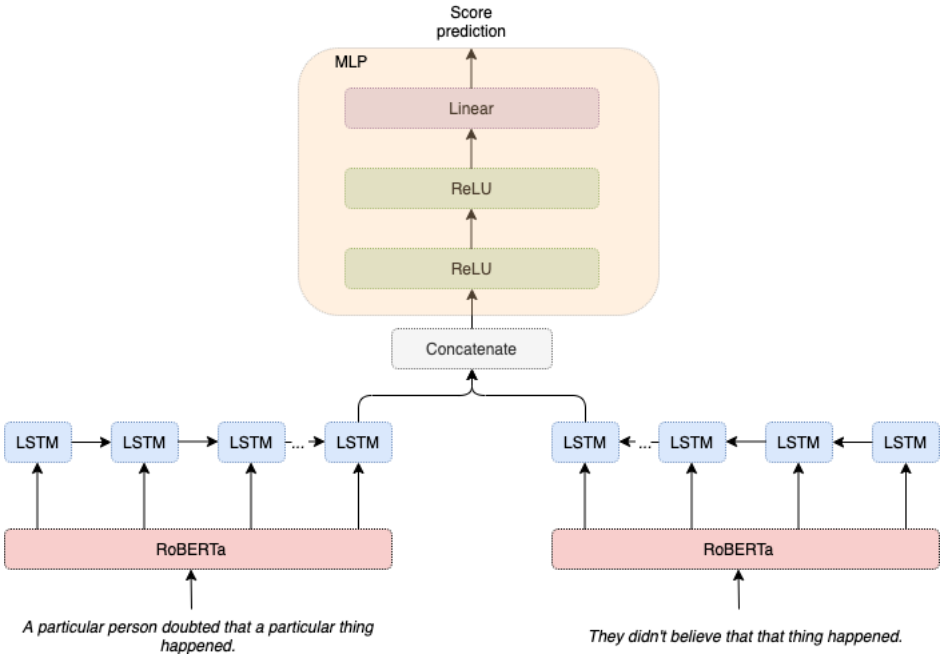

Figure 3: The LSTM+MLP model with RoBERTa embeddings used in our experiments. Note that this does not show the isotonic regression used for calibrating the network outputs.

optimize acceptability judgments since we do not collect those in the pilot. Therefore, for items with antecedent and consequent predicate labels p_a and p_c, antecedent and consequent negation labels n_a and n_c, and participant labels l, we optimize $\nu_{p_a p_c n_a n_c}$ against the following KL divergence loss:

$$\mathcal{L} = \sum_{l} D(r_{p_a p_c n_a n_c l} \parallel \hat{r}_{p_a p_c n_a n_c l}) \quad (1)$$

$$\hat{r}_{p_a p_c n_a n_c l} = \text{logit}^{-1}(m_l \nu_{p_a p_c n_a n_c} + \beta_0 + \beta_l)$$
$$\text{where } m_l = \exp(\sigma_0 + \sigma_l) \quad (2)$$

where σ_0 and β_0 are fixed scaling and shifting terms, and σ_l and β_l are random scaling and shifting terms for each participant l. The distribution of normalized ratings is shown in Figure 4.

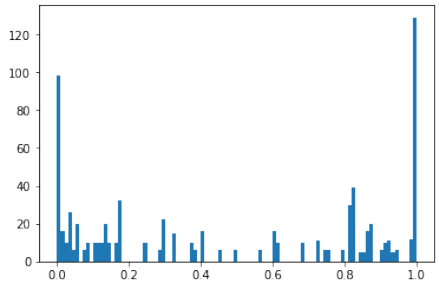

Figure 4: The distribution of normalized ratings (histogram).

3 Experiments and Results

3.1 Training and Evaluation

In our experiments, we trained the model described in §2.2 separately on three datasets—MegaVeridicality alone, MegaNegRaising alone, and the two together—and tested it on the pilot data.[4] For each dataset, the training and testing regime works as follows.

We first run the LSTM+MLP for up to 10 epochs over the entire dataset. Ideally, we would have used development data from the pilot to control training, but with only 96 examples, the dataset is simply too small to partition into dev and test sets of reasonable size. We used a batch size of 1 and Adam (Kingma and Ba, 2014) as our optimizer with a learning rate of 0.005 and $\beta_1 = 0.9$, $\beta_2 = 0.999$. We also apply a single dropout layer ($p = 0.5$) after concatenating the LSTM outputs and before the MLP.

Once training has finished, the network weights are frozen. We then split the pilot data into four folds of 24 examples. We pass all folds through the LSTM+MLP to obtain predictions and use the outputs for three of the folds to fit the isotonic regression. The network outputs for the fourth fold are used as the test set and are passed through the fitted regression to obtain *calibrated* predictions, against which we evaluate mean squared error. We repeat this procedure four times for each of our three datasets, using a different fold of the pilot data for testing each time.

With more data, we would certainly have run this on a GPU. However, given the relatively small size of all data sets ($\approx 8,000$ examples for MegaNegRaising and $\approx 5,000$ for MegaVeridicality), we found that training completed within just a couple of hours on a cluster of 24 2.40 GHz Intel Xeon CPUs with 62GB of memory.

[4] MegaVeridicality and MegaNegRaising both include annotator judgments about the grammatical acceptability of their sentences. We note that we did not filter out low acceptability examples, though this would be a good manipulation to attempt in future experiments.

3.2 Results

We report mean and standard deviation across folds for each of the three datasets in Table 1, as well as best train performance in Table 2.

We first observe that training loss for the MegaNegRaising model is significantly lower than for the other two. However, we have reason to believe this difference is somewhat artificial, and suspect that the result is more an artifact of our very small batch size than of any fundamental difficulties the model may have in learning the distributions of any of the datasets. Although the MegaNegRaising model did converge much faster than the other two (within a single epoch), the latte showed steady gains all the way through the 10^{th} epoch, and likely would have continued to do so had we allowed training to go on longer.

Second, we note the surprising *negative* R^2 values for all three datasets, though in no case is the value significantly different from 0. That the isotonic regression could do no better (and, in fact, seemed to do worse) than predicting the mean in all cases invites at least three possible explanations, not necessarily exclusive of one another:

(1) *The pilot dataset is simply too small or too noisy.* In addition to there being only 96 examples in the pilot, many of these annotations are not of the best quality. It therefore seems likely that the distribution of responses is itself too noisy to be able to model well.

(2) *Training on negation-raising or veridicality inferences is genuinely unhelpful in predicting the types of inferences in the pilot.* It's entirely plausible that the fea-

Dataset	mean	std
MegaNegRaising	-0.024	0.026
MegaVeridicality	-0.176	0.146
Combined	-0.182	0.189

Table 1: Average and standard deviation of the coefficient of determination (R^2) value across test folds.

Dataset	Loss
MegaNegRaising	0.015
MegaVeridicality	0.430
Combined	0.178

Table 2: Final training loss (mean squared error) for our model trained on MegaNegRaising alone, MegaVeridicality alone, and the two together.

tures the model learns that are relevant to determining whether a negation-raising or veridicality inference holds just aren't as relevant in this case. The fact that the pilot examples include two different verbs is another important consideration, and something that a model trained on sentences with only one verb may be less equipped to handle.

(3) *There is error in our normalization procedure.* Although we have spent a substantial amount of time scrutinizing the isotonic regression code, it is still a possibility that there are errors here that have escaped us.

The bulk annotation will allow us to rule out (1), and we intend to do more work to ensure that (3) is not the case. This will leave (2), which is our hypothesis, to be confirmed or invalidated.

4 Related Work

Our project draws primarily on two bodies of research: computational approaches to natural language inference and the literature on veridicality, factivity, and neg-raising, with a particular debt to the MegaAttitude project. We briefly discuss some of the relevant work from each domain below.

4.1 Natural Langauge Inference

Our model draws on an approach to natural language inference (NLI) that treats inference as a classification problem: given a premise text and a hypothesis text, predict whether the hypothesis is an entailment of the premise, a contradiction, or neither. To the best of our knowledge, the use of neural networks began with (Bowman et al., 2015), which introduced the Stanford Natural Language Inference (SNLI) dataset. The authors of that paper presented several different models, including an LSTM with MLP model in the vein of our own, though the MLP uses tanh activations instead of ReLUs and uses 300-dimensional GloVe embeddings for the inputs (Pennington et al., 2014).

Since the release of SNLI, there has been remarkable progress on NLI and on the related task of question answering (QA). Attention mechanisms, in particular—notably absent from the original SNLI models (and absent from our baseline implementation)—have found successful application in recent models. Some, such as ELMo (Peters et al., 2018) and its progeny, still rely on recurrent encoding architectures and use attention to focus on different subsets of intermediate representations within the encoding.

Increasingly, however, the most effective NLI models, such as SemBERT (Zhang et al., 2019), are based on the transformer architecture (Vaswani et al., 2017), which dispenses with RNNs altogether in the encoding in favor

of attention-only network layers. Both SemBERT and the RoBERTa embeddings that we rely on in our model are derived from the now-ubiquitous BERT language model (Devlin et al., 2018), which is transformer-based and which achieved state-of-the-art performance on both the SNLI and Stanford Question Answering (SQuAD) tasks when it was released.

One shortcoming of these tasks, however, is their failure to discriminate among different types of inference, though some effort has been made in the last several years to remedy this problem. Most notably, in (White et al., 2017), the authors "recast" several existing datasets for such tasks as paraphrase detection, anaphora resolution, and semantic role labeling, and probe how well a neural model trained on each one alone performs when evaluated on the others. A follow-up paper (White et al., 2018) extended this work to include nine additional datasets. We view our own project as forwarding this broader objective of more fine-grained analysis of the capabilities of modern NLI models.

4.2 MegaAttitude

MegaAttitude[5] is an effort to collect large-scale annotations of various semantic and lexical properties of English clause-embedding verbs, nouns, and adjectives, with the goal of better understanding their behavior. So far, the project includes four categories of annotation: acceptability (i.e. whether a particular frame is acceptable with a particular verb), veridicality, negation-raising, and orientation—how much the semantic interpretation of a phrase or sentence is due to the denotations of predicates and how much is due to their arguments.

We briefly highlight a couple of papers that were particularly relevant to our work. (An and White, 2019) extends a boolean matrix factorization model for S-selection initially presented in (White and Rawlins, 2016) to determine the extent to which a given neg-raising inference is explained by each of the following three factors:

(1) *Properties of the relation the verb denotes.* For example, *believe* denotes a certain relationship between a person and a proposition, and it's possible that some features of that relation explain why the neg-raising inference is licensed.

(2) *Properties of the kinds of things that the given verb can relate.* If *believe* relates people and propositions, it may be that certain facts about people or about propositions are what license the inference.

(3) *Whether the verb has a particular type signature.* In this context, a *type signature* is a latent variable that can be conceived as some unknown combination of syntactic and semantic features of the clause the variable takes. (This is treated as a hyperparameter in the model.)

They find that a latent variable model positing a single semantic and a single syntactic feature did the most to explain the neg-raising inferences. While we have been less interested in syntactic features in our pilot, we intend to explore their importance in our bulk study.

The MegaVeridicality dataset was initially presented in (White and Rawlins, 2018) and subsequently extended in (White et al., 2018). The original paper demonstrated that, contrary to existing opinions in the linguistics literature, there is virtually no correlation between the types of clauses a verb can select for and its tendency to license veridicality and factivity inferences when one considers a sufficiently large portion of English verbs. The second paper augments the first version of the dataset by adding additional clause types and

[5] http://megaattitude.io/

collecting more annotator judgments. Using this second version of the dataset, the authors probe several (then-)recent NLI models and find that they reliably make errors in veridicality judgments on certain types of sentences, most notably those involving a mismatch in polarity between clauses in the hypothesis and the antecedent.

5 Conclusion and Future Work

5.1 Discussion

The results presented here are, as mentioned, merely preliminary. Our future work will consist most immediately in the following:

(1) *Running and analyzing the second pilot.* and thereafter conducting a bulk data collection using all 800 of the verbs used in (An and White, 2019) to look more explicitly at semi-negation-raising inferences. Having less noisy data and more of it will enable more effective model training (as we will have enough data to do early stopping with a development set) and will also plainly allow us to better understand the inference patterns of different categories of verbs.

(2) *Exploring better model hyperparameters.* In particular, we are interested in exploring larger batch sizes, as we suspect that the slow convergence of the veridicality and neg-raising+veridicality models is due primarily to our purely stochastic (i.e. batch size 1) training regime.

(3) *Revisiting our normalization procedure.* We remain uncertain as to why the isotonic regression fails to properly calibrate the network outputs to the normalized true scores, and, as discussed in §3.2, this could be due to error on our part. It may also simply be preferable to handle normalization as a pre-processing step, possibly simply relying on z-scoring instead.

5.2 Individual Contributions

This report, the final presentation, and all the work behind it was a collaborative effort.

Ruoyang invested considerable energy in understanding the Amazon Mechanical Turk interface and shared her knowledge with Ben and Will. She was also heavily involved in developing and running the pilot and has taken charge of preparing the upcoming second pilot as well. She wrote the abstract and §1 of the report.

Ben worked extensively on the first pilot and, along with Will, went through numerous iterations of the code used to generate the CSV files for it. He has offered considerable input on the second pilot as well. Additionally, he helped to write and debug the code for the model, and to run the experiments. He wrote §2.1 and §2.3 of the report.

In addition to his work on the first pilot, Will wrote the majority of the code for the model, aided in debugging, and helped run the experiments. He wrote §2.2, §3.1, §3.2, §4, and §5.

5.3 Acknowledgments

This work was proposed and guided by Aaron White, to whom we express our deep gratitude. We would also like to thank Zhen Bai for leading the seminar that made the project possible and for offering feedback along the way; Hannah An, for providing code from her previous work; and the University of Rochester's Center for Integrated Research and Computing (CIRC) for providing computational resources.

References

Hannah Youngeun An and Aaron Steven White. 2019. The lexical and grammatical sources of neg-raising inferences. *arXiv:1908.05253 [cs]*, October. arXiv: 1908.05253.

Pranav Anand and Valentine Hacquard. 2014. Factivity, belief and discourse. *The art and craft of semantics: A festschrift for Irene Heim*, 1:69–90.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Michael Cohen. 2017. A note on belief, question embedding and neg-raising. In *International Workshop on Logic, Rationality and Interaction*, pages 648–652. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jon Robert Gajewski. 2005. *Neg-raising: Polarity and presupposition*. Ph.D. thesis, Massachusetts Institute of Technology.

Jon Robert Gajewski. 2007. Neg-raising and polarity. *Linguistics and Philosophy*, 30(3):289–328.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Nadine Theiler, Floris Roelofsen, and Maria Aloni. 2019. Picky predicates: why believe doesn't like interrogative complements, and other puzzles. *Natural Language Semantics*, 27(2):95–134.

Wataru Uegaki, Yasutada Sudo, et al. 2017. The anti-rogativity of non-veridical preferential predicates.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Aaron Steven White and Kyle Rawlins. 2016. A computational model of S-selection. *Semantics and Linguistic Theory*, 26(0):641–663, October.

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, USA. GLSA Publications*.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005.

Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2018. Lexicosyntactic inference in neural models. *arXiv preprint arXiv:1808.06232*.

Aaron Steven White. 2019a. Frequency, acceptability, and selection : A case study of clause-embedding.

Aaron Steven White. 2019b. Nothing's wrong with believing (or hoping) whether. *Under review at Semantics and Pragmatics*.

Aaron Steven White. 2020. Lexically triggered veridicality inferences. *Handbook of Pragmatics: 22nd Annual Installment*, 22:115.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2019. Semantics-aware bert for language understanding. *arXiv preprint arXiv:1909.02209*.