Predicting scalar inferences from "or" to "not both" using neural sentence encoders

Elissa Li
Stanford University
elissali@stanford.edu

Sebastian Schuster
Stanford University
sebschu@stanford.edu

Judith Degen
Stanford University
jdegen@stanford.edu

1 Introduction

Neural networks have recently successfully learned to predict some pragmatic inferences (e.g., Jeretic et al. (2020); Jiang and de Marneffe (2019)). For instance, Schuster et al. (2020) trained a neural network to predict human ratings of scalar inference strength from "some" to the negation of a stronger alternative with "all". However, it remains an open question to what extent these results are specific to the inference from "some" to "not all" or whether they generalize to other types of scalar inferences. We thus explore to what extent a neural network can learn to predict a different widely studied scalar inference: that from "or" to the negation of a stronger alternative with "and", as in (1).

(1) Jane **or** Alex came to the party. (inference: but they didn't both come)

Though "or" is typically treated as inclusive logical disjunction which can be pragmatically enriched to yield exclusivity (Chierchia et al., 2004; Sauerland, 2012), little is known about the natural distribution of "or" and the extent to which the scalar inference is context-dependent (but cf. Jasbi, 2018). Moreover, while a much smaller focus of the literature, "or" can take on many different pragmatic functions, including metalinguistic disjunction (Horn, 1985, *Let's go for a drink, ...or let's take a nap*), definitional uses (Potts and Levy, 2015, *A year has 12 months or 365 days*), and others. One estimate suggests there are at least 20 different readings of "or" (Ariel and Mauri, 2018). This poses two challenges, an empirical and a computational one. The empirical challenge is to establish the extent and type of context-dependence of scalar inferences from "or" to "not both". The computational challenge is to establish whether computational models can learn to predict context-dependent variability in inference strength.

We address these challenges as follows. In two web-based experiments in which participants rated the inference strength of naturally occurring utterances with "or", we first establish that there is considerable variability that can be partly explained by the presence of several linguistic and discourse features. We then use the data from these experiments to train neural network models to predict scalar inferences, and evaluate to what extent these networks learn to associate linguistic features with inference strength.

2 Empirical Challenge

We crowd-sourced inference strength ratings for 1,244 sentences with "or" from the Switchboard Corpus (Godfrey et al., 1992). Participants read 10-sentence paragraphs, where the last sentence of each paragraph was the target sentence with "or". Participants rated the similarity of the target to a comparison sentence which was identical to the target but included "but not both" concatenated to the end of the original disjunction. The crowdsourcing was run twice, and differed in the rating scale used. In the first iteration, similarity was rated on a sliding scale from 0–1 (*slider* dataset). In the second iteration, similarity was rated on a Likert scale from 1–7 (*Likert* dataset). In both iterations, the low endpoint was labeled "very different meaning" (indicating inclusive "or") and the high endpoint labeled "same meaning" (indicating exclusive "or"). Each sentence received 9 to 11 ratings in each dataset. The datasets were further annotated for a variety of theoretically motivated features that were expected to modulate inference strength. We focus on the presence of the lexical feature "either" (see (2)) and embedding of "or" in a downward-entailing context like negation (see (4)).

(2) You're either liberal or conservative.

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 446-450.
Held on-line February 14-19, 2021

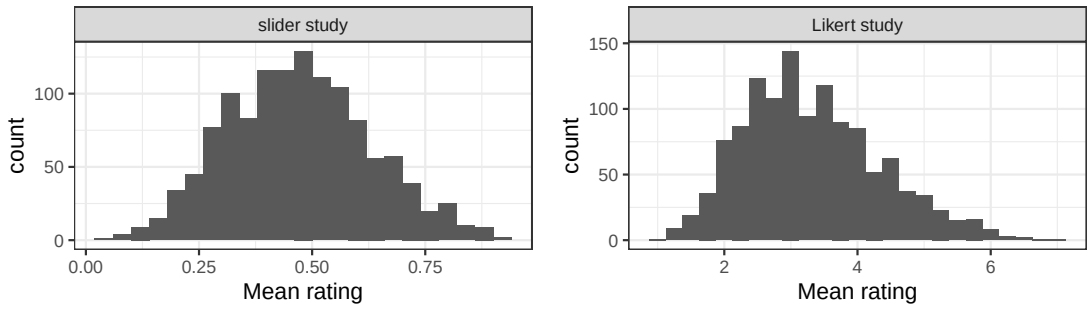

Figure 1: Distributions of mean ratings in *slider* dataset and *Likert* dataset.

(inference: but not both; means: slider=0.89, Likert=6.56)

(3) And it naturally, uh, composts or stores. (inference: but not both; slider=0.49; Likert=3.54)

(4) Well, I don't believe in drugs or alcohol. (inference: but not both; slider=0.15, Likert=1.11)

2.1 Results

We focus on reporting results on the Likert dataset, but the qualitative effects were identical across datasets. There was substantial variability in inference strength across items (see by-item means in Fig. 2). The response distributions for individual items often displayed bimodality, indicating disagreement among participants regarding whether or not the inference was intended. Mixed effects regressions revealed that the presence of "either" increased inference strength, and embedding under downward-entailing operators decreased inference strength (see model coefficients in dashed box of Fig. 2, original model). In comparison to the dataset reported for "some" (Degen, 2015), the overall variance captured by the current regression model was considerably lower (marginal $R^2 = 0.07$, conditional $R^2 = 0.24$), presumably in part due to the relative oddness of adding "but not both" to many of the naturally occurring sentences. This indicates that in contrast to a prevalent assumption in the literature, the sentence with "and" is not always a salient alternative to the sentence with "or", which suggests that alternative pragmatic functions of "or" are much more prevalent than the scarce attention they've received in the theoretical literature would suggest (but see Txurruka and Asher, 2008; Ariel and Mauri, 2018; Ariel, 2020).

3 Computational Challenge

Following Schuster et al. (2020), we experiment with a neural sentence encoder to predict the inference strength ratings. As a starting point, we use the same basic model,[1] which embeds the utterance using BERT (Devlin et al., 2019), passes the embeddings through a biLSTM layer followed by a self-attention layer.[2] While Schuster et al. (2020)'s model objective was to predict mean inference rating for each item, we additionally predict the distribution of inference ratings for each item, in order to better capture the higher participant disagreement and resulting bimodal rating distributions. To predict mean ratings, we minimize the mean squared error of the predicted ratings. To predict rating distributions, we explore three distribution types: 1) a Beta distribution, 2) a 2-component mixed Gaussian distribution, and 3) a discrete distribution of 7 buckets corresponding to the Likert scale ratings.[3] For distributions (1) and (2) we minimize the negative log-likelihood of the sample of human inference ratings; for distribution (3) we minimize the KL-divergence between the actual and predicted discrete distributions.

The model was trained on sentence-mean rating pairs for the first model objective (predicting mappings of sentences to mean inference ratings), and trained with sentence-distribution parameter pairs for the second model objective (mapping sentences to distribution of inference ratings). We

[1] We use the publicly available implementation from https://github.com/yuxingch/Implicature-Strength-Some.

[2] Alternatively, we could have also fine-tuned a pre-trained BERT model instead of training additional layers on top of BERT, which is more common-practice. We chose the latter approach to make our results better comparable to the results by Schuster et al. (2020)

[3] For the slider dataset, the continuous ratings were bucketed into 7 groups in order to replicate a 1–7 Likert scale.

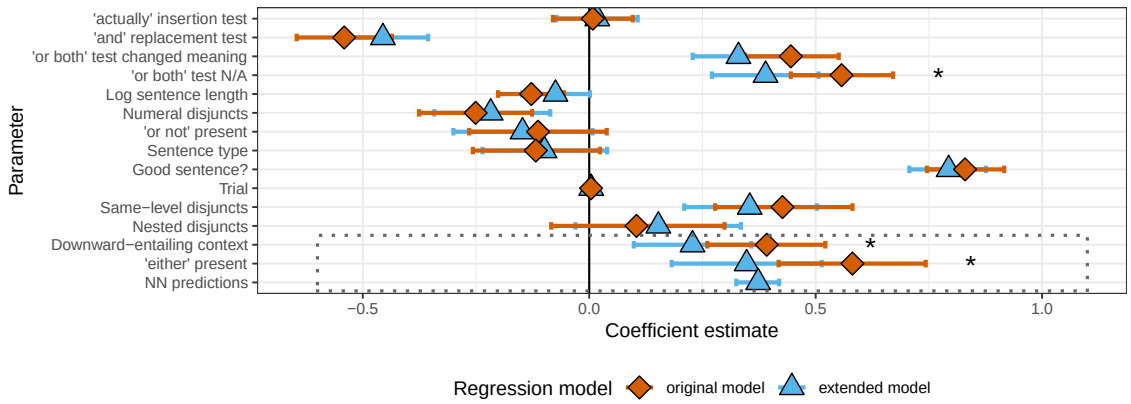

Figure 2: Mean estimates for regression coefficients and corresponding credible intervals for Bayesian mixed-effects model estimated from the Likert dataset. The *original model* coefficients were obtained from a model that does not include the neural network predictions as a predictor; the *extended model* contains all the predictors from the *original model* and the predictions from the neural network. (*: $p < 0.05$)

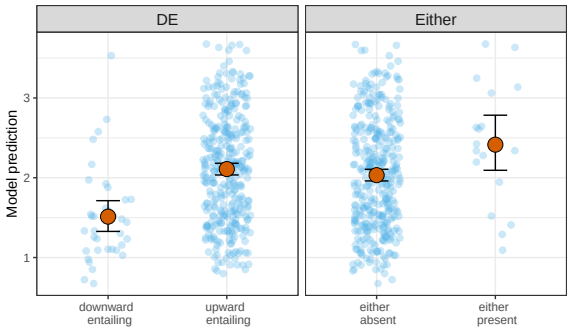

Figure 3: By-item and mean model predictions for utterances with "or" split by downward- or upward-entailing context (left panel) and split by the presence of "either" in the utterance (right panel).

use 5-fold cross-validation for hyperparameter tuning. Learning curves were also inspected during cross-validation to check for overfitting. Model performance was assessed by the correlation between the mean human inference ratings and the mean predicted rating.

3.1 Results

Cross-validated tuning showed the discrete model consistently outperformed the Beta and Gaussian models. We thus used the discrete model for further analyses. The best model's predictions correlated with the held-out test data ($r = 0.466$ on the Likert dataset, $r = 0.391$ on the slider dataset), but much less so than Schuster et al. (2020)'s model of inference strength in sentences with "some" (their $r = 0.78$).

3.2 Regression and Qualitative Analysis

To determine whether the model learned associations between lexical features and inference strength, we compared a Bayesian variant of the regression model described above to a regression model which also includes the network's output as a predictor (extended model). As Fig 2. shows, the coefficient for the downward entailment and 'either' present predictors is significantly smaller in the extended model, suggesting that the neural model captured most of the variance originally explained by these predictors. Further, a comparison of the neural model predictions on downward- vs upward-entailing items and items with and without "either" (Fig. 3) revealed that the neural model's predictions mirrored the effects identified by the regression model, further suggesting that the model learned the association between these cues and inference strength.

4 Discussion

The experimental and modeling results highlight several similarities between scalar inferences with "some" and scalar inferences with "or". 1) Both cases exhibit considerable contextual variability. 2) To some extent, this contextual variability is predicted by several linguistic and discourse features. 3) To some extent, neural sentence encoders can learn to predict the strength of the inference. 4) For both types of inferences, neural models are capable of learning associations between linguistic features and inference strength.

However, we also found important differences

between the two inference types. 1) There was considerable within-item variability in the experiments presented here. 2) The linguistic features discussed in the theoretical literature predict much less variance in the inference strength ratings of utterances with "or" than in the case of "some". 3) The neural network was worse at predicting inference strength for "or" than for "some."

Taken together, these findings suggest that predicting inference strength from surface cues is considerably more challenging in the case of "or".

We also found that predicting a discrete distribution led to better neural model performance than predicting a continuous distribution. This is somewhat surprising considering that (unlike the training objectives of the continuous distributions) the training objective of the discrete distribution does not consider the ordering of the different prediction values; predicting a rating of 1 instead of 7 leads to the same error magnitude as predicting a rating of 6 instead of 7. There are several potential explanations: One possibility is that training in the discrete model used a more standard training objective in the form of a KL-divergence loss (compared to the less common log likelihood objective). Noise in the empirical inference strength distributions is another possibility: the Beta and mixed Gaussian models both required fitting a distribution over the existing set of inference strength ratings for each sentence, and the resulting distributions may have been too noisy to learn.

Further, it is noteworthy that the model was only trained on the target utterances while human participants also had access to the preceding discourse context. This decision was based on the findings by Schuster et al. (2020) who found that incorporating the discourse context (at least in a naive way) did not improve prediction accuracy and therefore we limited the model's input to the target utterance. Developing more sophisticated methods to integrate the larger conversational context remains an important avenue for future work, given that humans incorporate the discourse context when drawing inferences.

Finally, as mentioned in the introduction, Jeretic et al. (2020) recently also evaluated the ability of neural network models to draw pragmatic inferences based on an automatically constructed evaluation dataset derived from a grammar. However, their work was primarily aimed at evaluating whether pre-trained natural language inference models consistently make predictions using the logical meaning of "or" or whether they consistently use the pragmatic meaning of "or". Because of that, their evaluation dataset is built with the assumption that an occurrence of "or" always triggers the inference to "not both". Thus, while their work–similarly to ours–was concerned with neural network model's abilities to draw pragmatic inferences, their and our results are not easily comparable since their evaluation was not concerned with any form of variability in naturally occurring examples.

5 Conclusion

We have shown in this work that neural networks can to some extent learn to predict inferences from "or" to "not both," but this inference appears more complex than other scalar inferences. Further investigation into the human factors influencing inference strength as well as more sophisticated models are necessary to fully explain the phenomenon of "or" interpretation.

Acknowledgements

We gratefully acknowledge Leyla Kursat for implementing and running the web-based experiments. This material is based upon work supported by the National Science Foundation under Grant #2030859 to the Computing Research Association for the CIFellows Project.

References

Mira Ariel. 2020. Or constructions, argumentative direction and disappearing 'alternativity'. *Language Sciences*, 81:101195. Ad hoc categorization and language: the construction of categories in discourse.

Mira Ariel and Caterina Mauri. 2018. Why use or? *Linguistics*, 56(5):939–993.

Gennaro Chierchia et al. 2004. Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Structures and beyond*, 3:39–103.

Judith Degen. 2015. Investigating the distribution of 'some' (but not 'all') implicatures using corpora and web-based methods. *Semantics & Pragmatics*, 8(11):1–55.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*, pages 4171–4186.

John J. Godfrey, Edward C. Holliman, and Jane. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 92)*.

Laurence Horn. 1985. Metalinguistic negation and pragmatic ambiguity. *Language*, 61(1):121–174.

Masoud Jasbi. 2018. *Learning Disjunction*. Ph.D. thesis, Stanford University.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Do you know that florence is packed with visitors? evaluating state-of-the-art models of speaker commitment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*.

Christopher Potts and Roger Levy. 2015. Negotiating lexical uncertainty and speaker expertise with disjunction. In *Proceedings of the 41st Annual Meeting of the Berkeley Linguistics Society*.

Uli Sauerland. 2012. The Computation of Scalar Implicatures: Pragmatic, Lexical or Grammatical? *Language and Linguistics Compass*, 6(1):36–49.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*.

Isabel Gomez Txurruka and Nicholas Asher. 2008. A discourse-based approach to natural language disjunction (revisited). In *Language, Representation and Reasoning*. University of the Basque country Press.