Inclusivity-Exclusivity Inference Using Recurrent Neural Networks

Stanford CS224N Custom Project. Mentors: Hang Jiang, Judith Degen (external)

OPTION 3

Elissa Li
elissali@stanford.edu

Santosh Murugan
smurugan@stanford.edu

Abstract

Though “or” in formal semantics and logic is often understood as a logical operator with a well-defined truth table, the exclusivity vs. inclusivity of “or” in natural language often fluctuates sentence-to-sentence and depends on subtle linguistic cues. This paper explores the ability of neural networks to predict the behavior of “or” across various sentence structures. We assess the performance of a biLSTM-based sentence encoder trained on an English dataset of human inference inclusivity-exclusivity ratings, experimenting with three different pre-trained word embedding models, with and without self-attention. We find the best-performing model to utilize BERT embeddings without attention, which exceeds expectations by predicting human inference ratings with $r = 0.35$.

1 Introduction

The word “or” is often thought of in its capacity as a logical connective (e.g. p v q) – to express an inclusive or exclusive choice between two options. In natural language, however, “or” tends to be less binary, and can function in up to eight different manners [1]. Often, the inclusivity-exclusivity judgments that English-language speakers make are based on intuition and linguistic experience, which further complicates interpretation by listeners. The question of how to determine whether and to what extent “or” is meant inclusively or exclusively in a given context is an interesting linguistic challenge, and can additionally help identify specific sentence-level features that are strong indicators of inclusivity or exclusivity.

In recent years, several studies have focused on predicting scalar inference for similar utterances via classical methods such as Bayesian inference and RSA models of language understanding [2] [3] [4]. While these methods have drawbacks (in some cases, requiring manual feature engineering - see Section 2.1), neural network models have shown promise in predicting inference strength for related tasks [5], indicating that sentences likely contain linguistic cues which could aid interpretation.

In this study, we analyze the extent to which neural network models can predict the behavior of “or” in various sentence structures. Building off the work of [5], who model the behavior of "some" with regard to scalar inference strength, we apply several bidirectional LSTM-based model architectures to an English dataset, experimenting with three types of word embeddings: GloVe, BERT, and BERT-Large. We then adjust model complexity to optimize performance via hyperparameter tuning, and predict inclusive and exclusive inferences. After testing, we perform both quantitative and qualitative analyses of the results, including identifying sentence-specific features which may hinder models’ predictive power.

We find that using a bidirectional LSTM without self-attention and BERT embeddings outperforms the other embedding/attention combinations, yielding a correlation coefficient of 0.35 between the model and human predictions, which exceeds expectations. Furthermore, during the qualitative analysis, we notice that certain aspects of sentence structure (e.g. number of words, placement of "or" within the sentence) increase the difficulty of the predictive task. Based on these findings, we conclude that, while this model represents a substantial effort to model inference strengths, further improvements are required to capture the high pragmatic complexity of "or." Several potential avenues for future work are discussed.

2 Related Work

Computational modeling of scalar inference judgments have only recently moved towards neural network models, having historically centered around Bayesian statistical modeling. Furthermore, there has been no neural network modeling of scalar inference judgments with respect to "or" specifically. Below, we first describe recent work in scalar inference modeling. Next, we describe relevant linguistic research around the pragmatic behavior of "or" and its implications for computational modeling.

2.1 Previous Work in Computational Modeling of Scalar Inference

Previous work in predicting scalar inference has relied on Bayesian game-theoretic models of pragmatic reasoning (i.e. using Bayesian inference to recover speakers' intended meaning). Popular models include the rational speech act (RSA) model of language understanding, which aims to formalize the social inference view of pragmatics by modeling communication as a signalling game between the speaker and listener with Bayesian statistics as described in [2]. For instance, [3] uses RSA modeling to predict the scalar inference of "some" and find a good fit between model predictions and human judgments. [4] similarly uses RSA to model speaker uncertainty with respect to scalar inference judgments of "some" and finds the model successfully predicts listener interpretations of utterances.

However, though Bayesian models are successful in using speaker expectations to predict pragmatic inferences, they are limited in that they require manual specification of salient linguistic cues and specification of a finite set of possible inferences. For that reason, we refer to [5] for our predictive task. The authors proposed an LSTM-based sentence encoder as an alternative model of pragmatic reasoning, as neural network models do not suffer from these same limitations – they are able to make predictions for arbitrary utterances and do not require manual feature specification. The authors found that their model successfully learned linguistic features to predict subtle differences in scalar inferences with high accuracy ($r = 0.78$).

As previous neural network modeling of scalar inference has mostly been limited to the pragmatic behavior of "some," our task is a logical next step in evaluating the broader applicability of neural network models to pragmatic inference predictions. In expanding the project to "or," we reference both the performance benchmarks in [5] for "some" and the linguistic analysis of the pragmatic behavior of "or" in Section 2.2 below to inform our expected baseline performance.

2.2 Pragmatic Judgments of "Or" Utterances

Though "or" is often treated as a simple logical disjunction which can be pragmatically enriched to yield exclusivity vs. inclusivity, linguistic literature suggests that "or" in fact takes on as many as six additional pragmatic functions, yielding a total of seven distinct pragmatic roles. Besides (1) the logical/inclusive function and (2) the exclusive function, the following six pragmatic roles have also been distinguished for "or":

(3) a correction or "metalinguistic disjunction," in which the second disjunct replaces the first [6] – e.g. "He likes cake, or at least he likes sweet things";

(4) an equivalence relation wherein two synonymous terms are conjoined, usually with one unfamiliar term being clarified by a relatively more common one [7] – an example from the corpus: "Sole or flounder is real good";

(5) an imperative [8], e.g. "Sit down or else!";

(6) a conjunctive [8], e.g. "They like to be able to attract the Einsteins, or the Professor Chou. . . ";

(7) a cue for uncertainty [8], e.g. "Oh I don't know. She went to the movies or had a drink with her friends"; and

(8) preference [8], e.g. "Let's go for a drink or. . . let's take a nap."

The diversity of pragmatic functions that "or" can take on presents a challenge to computational models seeking to predict its inference judgments. We seek to evaluate the applicability of neural network modeling towards this more challenging task relative to historical work done for "some,"

with the intention of contributing towards a more general picture of the feasibility of neural networks to predict pragmatic judgments.

3 Approach

We base our approach on the model proposed by Schuster et. al. (2019) [5] in their study of scalar inference. In their tradition, the goal was to predict the mean inference rating of a sequence of words. We embed the words using pre-trained embeddings - selecting between GLoVE, BERT, or BERT-large- which are then passed through a bidirectional LSTM, with outputs mapped to a score between 0 and 1. Self-attention and dropout functionality are also implemented with usage varying across experiments.

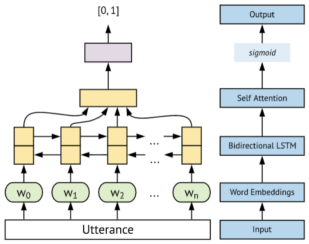

Figure 1: Model architecture [5]

The baseline model uses 100-dimensional GloVe for the word embedding model and does not include the self-attention layer; the model directly passes the output representation from the biLSTM through the sigmoid activation to obtain the scaled outputs.

We compare the performance of the baseline model (GloVe without attention) with four additional variants: GloVe with attention, BERT with and without attention, and BERT-large with and without attention.

To that end, we repurpose the model code from [5] [9] to accommodate our dataset. We begin by modifying the data preprocessing pipeline. The original model code considers linguistic cues not relevant to the "or" task, such as the presence of partitives, and did not perform optimal sentence preprocessing. We write scripts to preprocess each sentence, truncating target utterances, and transformed the dataset to have three features: a unique example identifier, preprocessed sentence, and mean rating.

In the post-processing phase, we additionally modify the source implementation to drop duplicate sentences, and to create separate directories to store train/test data and results. For each of the model variants, we perform hyperparameter tuning (modifying the learning rate, number of layers, maximum sequence length, number of hidden dimensions, dropout, and cross-validation) and assess performance.

4 Experiments

Data: The dataset was collected by Degen (2015) [10] and consists of 1,244 unique sentences. Participants from Amazon's Mechanical Turk were given paragraphs with ten sentences each. The last sentence (e.g. sentence (1a)) of each paragraph was highlighted and featured a single disjunction. Participants were then given a comparison sentence (1b), which was identical to (1a) with the addition of "but not both" concatenated to the end of the original disjunction:

1. (a) So I like things like Golden Girls or Cheers.
   (b) So I like things like Golden Girls or Cheers but not both.

Participants were asked to rate how similar in meaning the comparison sentence was to the original sentence on a sliding scale of 0-1, with 0 being completely different (inclusive *or*) and 1 being the same (exclusive *or*).

Evaluation method: We assess the performance of the baseline model and its improved variants by evaluating the correlation coefficient r of its predictions with the actual mean human inference ratings for each utterance. Specifically, we use 5-fold cross validation, and average the r across across each fold to get the mean validation r. Lastly, for the best performing model from each embedding type, we evaluate performance on a held-out test set consisting of 372 sentences (30%) from the original data.

We evaluate our performance relative to the performance benchmarks in [5] for "some" and expect our baseline performance for the prediction of "or" to be significantly worse than that of the "some" paper. This is because we expect the "or" task to be significantly harder: Whereas "some" serves very limited linguistic functions, "or" plays as many as 8 different discourse functions, some of which include: logical/inclusive ("do you want something to eat or drink"), exclusive ("either x or y"), corrections [6], equivalence [7], and imperative [8]. The dataset reflects this higher complexity (Figure 2). Whereas the "some" dataset saw more extreme judgments of inference strength (indicating high human certainty in their judgments) resulting in a more bimodal distribution, the "or" dataset saw a more unimodal distribution with ratings clustered near 0.5, indicating high human uncertainty in the correct interpretation of "or". Given higher human uncertainty, we also expect worse performance and lower correlation of our model's predictions to human inference ratings.

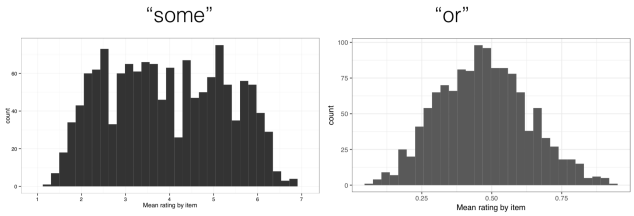

Figure 2: Distribution of inference rating in dataset [1]

We therefore define a "satisfactory" performance to be an r of 0.2 - 0.3, due to the relatively higher complexity of the linguistic task and higher human uncertainty reflected in the dataset.

Experimental details: After the data preprocessing described in "Approaches," we ran the previously described six model variants (3 embeddings, each with and without attention).

The LSTM models are implemented in PyTorch, and we use 5-fold cross-validation on the training data to optimize the following hyperparameters, comparing all combinations of each parameter (see Figure 3): Word embedding model: 100d GloVe, 768d BERT-base; and Dropout rate in LSTM: 0, 0.3, 0.5. For each model we maintain a hidden layer size of 100 neurons. Training is optimized with Adam, using a learning rate of 0.001 (with a decay rate of 0.8 starting at epoch 20) to minimize the mean-squared-error loss function. All models are trained on CPU (no CUDA) for 800 epochs. Training time was 33 minutes per fold.

5 Results and Analysis

						Validation Set	Test Set	
EMBEDDING	Layers	Dimension	Dropout	LR	Attention	best avg_r	best_r	best_epoch
BERT	2	100	0.5	0.001	FALSE	0.349	0.345	500
BERT	1	100	0	0.001	TRUE	0.347		
BERT	2	100	0.3	0.001	FALSE	0.347		
BERT	1	100	0	0.001	FALSE	0.343		
BERT Large	2	100	0.3	0.001	TRUE	0.323	0.302	210
BERT Large	2	100	0.5	0.001	FALSE	0.321		
BERT Large	2	100	0.5	0.001	TRUE	0.321		
BERT	2	100	0.3	0.001	TRUE	0.316		
BERT Large	2	100	0.3	0.001	FALSE	0.315		
BERT	2	100	0.5	0.001	TRUE	0.315		
BERT Large	1	100	0	0.001	TRUE	0.314		
BERT Large	1	100	0	0.001	FALSE	0.309		
GloVe	2	100	0.5	0.001	FALSE	0.301	0.279	10
GloVe	2	100	0.3	0.001	FALSE	0.296		
GloVe	2	100	0.5	0.001	TRUE	0.282		
GloVe	1	100	0	0.001	FALSE	0.281		
GloVe	2	100	0.3	0.001	TRUE	0.280		
GloVe	1	100	0	0.001	TRUE	0.280		

Figure 3: Hyperparameter Tuning Combinations

Figure 3 provides a summary of each model and hyperparameter combination that was implemented. As seen above, the best validation r is approx. 0.35, which is well above the expected r = 0.2 - 0.3. Analyzing the data further, we see that BERT without attention strictly outperforms all other models, while GloVe performs strictly worse than all other models. The relatively lower performance of GloVe was expected and mimics the results from Schuster et al.

We attribute the superior performance of BERT versus BERT-large to the possibility that BERT-large (as an inherently more complex embedding style) may have been more prone to overfitting, resulting in lower validation and test r's. The same argument regarding overfitting can be made for why including self-attention tended to hurt performance. Although we implemented various techniques to combat overfitting (adjusting learning rate via Adam, decreasing the number of layers, changing dropout rate, and changing the number of layers), performance did not increase - suggesting potential for additional hyperparameter tuning in future research.

5.1 Regression Plots - Distribution of Predictions

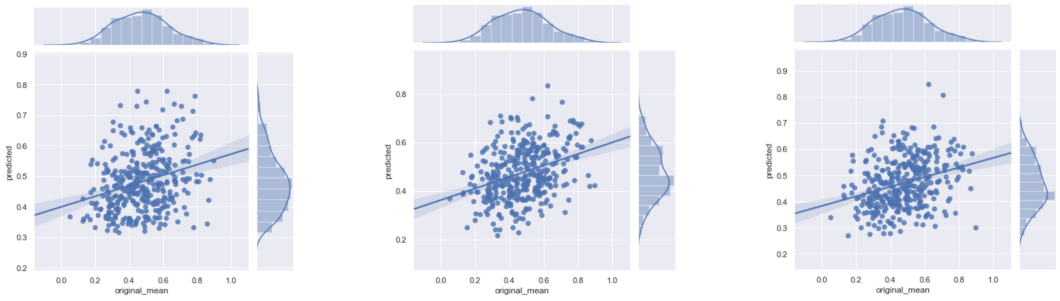

Figure 1: GloVe, BERT, and BERT-Large Ratings Distributions

To better visualize the correlations between our model and human predictions, for each of our three embeddings, we map the mean human ratings of "or"-exclusivity for each sentence against the model's prediction. We also plot the distribution of human ratings (top bar graph) and the distribution of the model's predicted ratings (sideways bar graph) for each.

In the case of GloVe (left), we observe a correlation of 0.28, and see that the distribution of predictions skews left/lower (< 0.5) than the distribution of human ratings. For BERT (center) and BERT-Large (right), we observe correlations of 0.35 and 0.30 respectively, and note that, while the distribution of predicted ratings skews slightly more left than the distribution of actual ratings, they generally appear to capture the shape of the distribution of human ratings.

5.2 Regression Plots - Macroscopic Changes in Correlation during Training

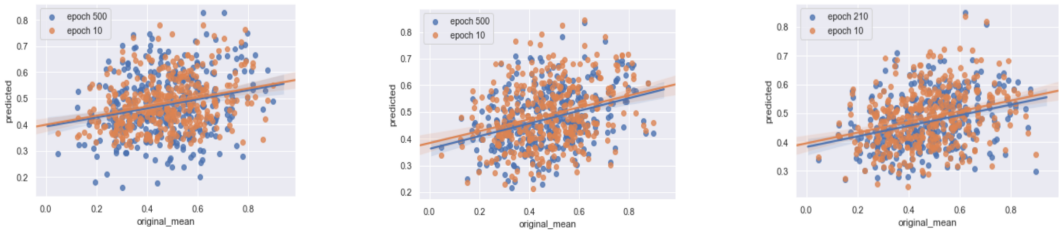

Figure 2: GloVe, BERT, and BERT-Large Ratings from Epoch 10 -> Best Epoch

Continuing with the visualization process, we map the scatter plots of model prediction correlation against human ratings at two epochs: epoch 10 (representing the early stages of training), and the best-performing epoch (with the highest correlation) for each model. The highest-correlation epochs are 10, 500 and 210, for GloVe, BERT, and BERT-Large respectively. These values are naturally different because each embedding type learns idiosyncratically; but in each case, the correlations increase until they reach an optimal value over time, indicating that the model is training correctly.

5.3 Regression Plots - Movement of Individual Data Points During Training

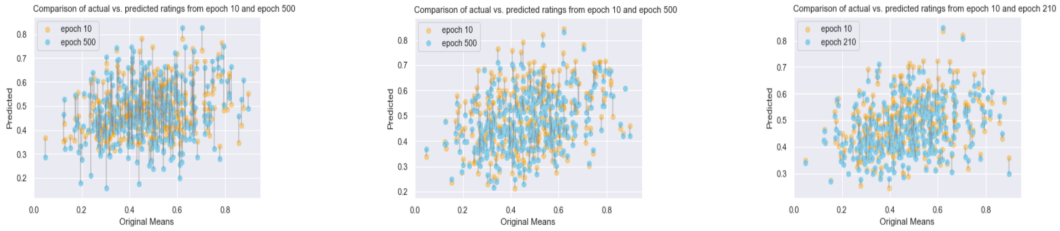

Figure 3: GloVe, BERT, and BERT-Large Ratings from Epoch 10 -> Best Epoch

The lollipop scatterplots in Figure 3 above trace the movement of each datapoint from epoch 10 to the highest-correlation epoch for each model, as mentioned in section 4.2.

For GloVe, we see that more centered predictions from epoch 10 tend to get pulled downwards by later epochs (i.e. notice that the datapoints on the lower outskirts of the scatterplot are all blue dots, i.e. from epoch 500). This suggests that the model errs on the side of lower inclusivity-exclusivity ratings over time.

For BERT and BERT-Large, we see that more extreme (and incorrect) predictions from epoch 10 tend to get reigned in by later epochs to be closer to center (e.g. notice that the points on the upper outskirts of the scatter plot tend to be from epoch 10, and are pulled closer to center by epoch 210).

5.4 Qualitative Analysis

For each of the embedding types, we sorted and examined individual sentences from the corpus that produced the highest differences between predicted and human ratings. For all three embeddings, the most clear trend was that each of the most highly-missed sentences was either significantly longer or shorter (in many cases by more than 50%) than the mean sentence length, which was approximately 24 tokens. This observation may be explained by the fact that whereas longer sentences may present too much linguistic noise, shorter sentences conversely tend to rely more on world knowledge (e.g. recognizing "half empty or half full" as presenting two mutually exclusive disjuncts) given the relative lack of linguistic cues.

Moreover, in each of these sentences, the placement of "but not both" (which was removed in the sentence pre-processing phase) was almost always towards the very end of the sentence. This might indicate that the models might be focusing too much on noise earlier in the sentence, hindering the prediction of exclusivity at the end of the sentence.

With regard to the eight linguistic functions of "or" outlined in [1] (inclusive, exclusive, corrections, equivalency, imperative, conjunction, uncertainty, and preference), the latter six do not appear to be at cause for significant model error. However, it is worth noting that the models tend to miss on cases where the human ratings are quite extreme (i.e. cases in which the "or" is heavily inclusive or heavily exclusive). In these cases, the models tend to make predictions closer to 0.5, indicating that the models have not yet learned to pick up on cues for extreme inclusivity or exclusivity.

Lastly, four out of the top five highest-missed sentences by BERT-Large were also missed by BERT, GloVe, or both, indicating that these particular sentences might have additional cues (other than abnormal sentence length and placement of "but not both") that make rating prediction challenging.

6 Conclusions and Future Work

In this work, we examine the ability of biLSTM-based sentence encoders to predict human inclusivity-exclusivity inferences. We optimize for hyperparameters and experiment with three different pre-trained embeddings (GloVe, BERT, and BERT-Large) as well as attention, testing the best performing model for each pre-trained embedding. We find that BERT embeddings with high dropout and without attention perform best, albeit still with an accuracy significantly worse than that of the “some” baseline (r = 0.35), as expected, owing to the increased complexity of the predictive task. From our quantitative and qualitative results, we conclude that the current model is suboptimal for utterances

with high pragmatic complexity like "or," and therefore not yet generalizable to scalar inference judgments broadly beyond the scope of "some."

Note that this conclusion is tempered by the high uncertainty and variance in human inference judgments reflected in our dataset. Future work towards confirming our results would begin with re-collecting data on the same utterances with a new set of participants to assess how predictive the first dataset is of the second, i.e. how consistent human inference judgments are for each of the test sentences across multiple trials.

Future avenues of research would investigate the impact of the presence of specific syntactic, semantic, and pragmatic features on prediction correlations, for example by identifying clusters based on the distinct linguistic functions of "or" using unsupervised learning and observing whether the model better learns certain linguistic functions over others. Future work could utilize these observations in better harnessing these relevant linguistic features to maximize prediction correlations.

References

[1] Judith Degen. On the natural distribution of 'some' and 'or': consequences for theories of pragmatic inference. In *Inaugural Colloquium SFB 1102, Phase II Universitat des Saarlandes*, 2019.

[2] Noah Goodman and Michael Frank. Pragmatic language interpretation as probabilistic inference. In *Trends in Cognitive Sciences, Vol 20, Issue 11*, 2016.

[3] Noah Goodman and Andreas Stuhlmuller. Knowledge and implicature: Modeling language understanding as social cognition. In *Topics in Cognitive Sciences, Vol 5, Issue 1*, 2013.

[4] Degen Judith, Michael Henry Tessler, and Noah Goodman. Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society (CogSci 2015), pages 548-553*, 2015.

[5] Sebastian Schuster, Yuxing Chen, and Judith Degen. Harnessing the linguistic signal to predict scalar inferences. In *Association for Computational Linguistics (ACL)*, submitted for review.

[6] Laurence Horn. Metalinguistic negation and pragmatic ambiguity. In *Language, Vol. 61, No. 1*, 1985.

[7] Catherine Ball. Metalinguistic disjunction. In *Penn Review of Linguistics*, 1986.

[8] Isabel Gomez Txurruka and Nicholas Asher. A discourse-based approach to natural language disjunction (revisited). In *M. Aunargue, K. Korta and J. Lazzarabal (eds.) Language, Representation and Reasoning, University of the Basque country Press*, 2008.

[9] Yuxing Chen. Harnessing the linguistic signal to predict scalar inferences, Github repository. `https://github.com/yuxingch/Implicature-Strength-Some`.

[10] Judith Degen. Investigating the distribution of 'some' (but not 'all') implicatures using corpora and web-based methods. In *Semantics Pragmatics, Volume 8, Article 11*, 2015.