Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition

Paloma Jeretič[*1], **Alex Warstadt**[*1], **Suvrat Bhooshan**[2], **Adina Williams**[2]
[1]Department of Linguistics, New York University
[2]Facebook AI Research
{paloma,warstadt}@nyu.edu, {sbh,adinawilliams}@fb.com

Abstract

Natural language inference (NLI) is an increasingly important task for natural language understanding, which requires one to infer whether a sentence entails another. However, the ability of NLI models to make pragmatic inferences remains understudied. We create an IMPlicature and PRESupposition diagnostic dataset (IMPPRES), consisting of >25k semi-automatically generated sentence pairs illustrating well-studied pragmatic inference types. We use IMPPRES to evaluate whether BERT, InferSent, and BOW NLI models trained on MultiNLI (Williams et al., 2018) learn to make pragmatic inferences. Although MultiNLI appears to contain very few pairs illustrating these inference types, we find that BERT learns to draw pragmatic inferences. It reliably treats scalar implicatures triggered by "some" as entailments. For some presupposition triggers like *only*, BERT reliably recognizes the presupposition as an entailment, even when the trigger is embedded under an entailment canceling operator like negation. BOW and InferSent show weaker evidence of pragmatic reasoning. We conclude that NLI training encourages models to learn some, but not all, pragmatic inferences.

1 Introduction

One of the most foundational semantic discoveries is that systematic rules govern the inferential relationships between pairs of natural language sentences (Aristotle, *De Interpretatione*, Ch. 6). In natural language processing, Natural Language Inference (NLI)—a task whereby a system determines whether a pair of sentences instantiates in an entailment, a contradiction, or a neutral relation—has been useful for training and evaluating models on sentential reasoning. However, linguists and philosophers now recognize that there are separate semantic and pragmatic modes of reasoning (Grice, 1975; Clark, 1996; Beaver, 1997; Horn and Ward, 2004; Potts, 2015), and it is not clear which of these modes, if either, NLI models learn. We investigate two pragmatic inference types that are known to differ from classical entailment: scalar implicatures and presuppositions. As shown in Figure 1, implicatures differ from entailments in that they can be denied, and presuppositions differ from entailments in that they are not canceled when placed in entailment-cancelling environments (e.g., negation, questions).

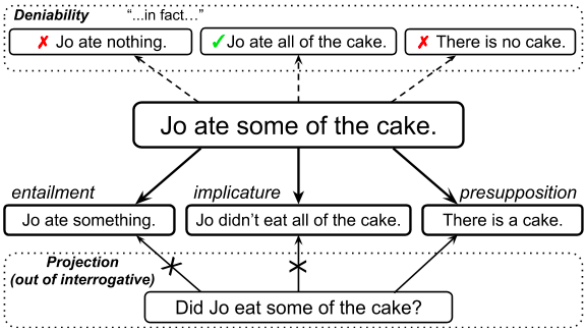

Figure 1: Illustration of key properties of classical entailments, implicatures, and presuppositions. Solid arrows indicate valid commonsense entailments, and arrows with X's indicate lack of entailment. Dashed arrows indicate follow up statements with the addition of *in fact*, which can either be acceptable (marked with '✗') or unacceptable (marked with '✓').

To enable research into the relationship between NLI and pragmatic reasoning, we introduce IMPPRES, a fine-grained NLI-style diagnostic test dataset for probing how well NLI models perform implicature and presupposition. Containing 25.5K sentence pairs illustrating key properties of these pragmatic inference types, IMPPRES is automatically generated according to linguist-crafted templates, allowing us to create a large, lexically varied, and well controlled dataset targeting specific

[*]Equal Contribution

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8690–8705
July 5 - 10, 2020.

instances of both types.

We first investigate whether presuppositions and implicatures are present in NLI models' training data. We take MultiNLI (Williams et al., 2018) as a case study, and find it has few instances of pragmatic inference, and almost none that arise from specific lexical triggers (see §4). Given this, we ask whether training on MultiNLI is sufficient for models to *generalize* about these largely absent commonsense reasoning types. We find that generalization is possible: the BERT NLI model shows evidence of pragmatic reasoning when tested on the implicature from *some* to *not all*, and the presuppositions of certain triggers (*only*, cleft existence, possessive existence, questions). We obtain some negative results, that suggest that models like BERT still lack a sophisticated enough understanding of the meanings of the lexical triggers for implicature and presupposition (e.g., BERT treats several word pairs as synonyms, e.g., most notably, *or* and *and*).

Our contributions are: (i) we provide a new diagnostic test set to probe for pragmatic inferences, complete with linguistic controls, (ii) to our knowledge, we present the first work evaluating deep NLI models on specific pragmatic inferences, and (iii) we show that BERT models can perform some types of pragmatic reasoning very well, even when trained on NLI data containing very few explicit examples of pragmatic reasoning. We publicly release all IMPPRES data, models evaluated, annotations of MultiNLI, and the scripts used to process data.[1]

2 Background: Pragmatic Inference

We take pragmatic inference to be a relation between two sentences relying on the utterance context and the conversational goals of interlocutors. Pragmatic inference contrasts with semantic entailment, which instead captures the logical relationship between isolated sentence meanings (Grice, 1975; Stalnaker, 1974). We present **implicature** and **presupposition** inferences below.

2.1 Implicature

Broadly speaking, implicatures contrast with entailments in that they are inferences suggested by the speaker's utterance, but not included in its literal (Grice, 1975). Although there are many types of implicatures we focus here on **scalar implicatures**. Scalar implicatures are inferences, often optional,[2] which can be drawn when one member of a memorized lexical scale (e.g., ⟨*some, all*⟩) is uttered (see §6.1). For example, when someone utters *Jo ate some of the cake*, they suggest that *Jo didn't eat all of the cake*, (see Figure 1 for more examples). According to Neo-Gricean pragmatic theory (Horn, 1989; Levinson, 2000), the inference *Jo didn't eat all of the cake* arises because *some* has a more informative lexical alternative *all* that could have been uttered instead. We expect the speaker to make the most informative true statement:[3] as a result, the listener should infer that a stronger statement, where *some* is replaced by *all*, is false.

Type	Example
Trigger	Jo's cat yawned.
Presupposition	Jo has a cat.
Negated Trigger	Jo's cat didn't yawn.
Modal Trigger	It's possible that Jo's cat yawned.
Interrog. Trigger	Did Jo's cat yawn?
Cond. Trigger	If Jo's cat yawned, it's OK.
Negated Prsp.	Jo doesn't have a cat.
Neutral Prsp.	Amy has a cat.

Table 1: Sample generated presupposition paradigm. Examples adapted from the 'change-of-state' dataset.

Implicatures differ from entailments (and, as we will see, presuppositions; see Figure 1) in that they are **deniable**, i.e., they can be explicitly negated without resulting in a contradiction. For example, someone can utter *Jo ate some of the cake*, followed by *In fact, Jo ate all of it*. In this case, the implicature (i.e., *Jo didn't eat all the cake* from above) has been denied. We thus distinguish implicated meaning from literal, or logical, meaning.

2.2 Presupposition

Presuppositions of a sentence are facts that the speaker takes for granted when uttering a sentence (Stalnaker, 1974; Beaver, 1997). Presuppositions are generally associated with the presence of certain expressions, known as **presupposition triggers**. For example, in Figure 1, the definite de-

[1] github.com/facebookresearch/ImpPres

[2] Implicature computation can depend on the cooperativity of the speakers, or on any aspect of the context of utterance (lexical, syntactic, semantic/pragmatic, discourse). See Degen (2015) for a study of the high variability of implicature computation, and the factors responsible for it.

[3] This follows if we assume that speakers are cooperative (Grice, 1975) and knowledgeable (Gazdar, 1979).

scription *the cake* triggers the presupposition that there is a cake (Russell, 1905). Other examples of presupposition triggers are shown in Table 1.

Presuppositions differ from other inference types in that they generally **project** out of operators like questions and negation, meaning that they remain valid inferences even when **embedded under** these operators (Karttunen, 1973). The inference that there is a cake survives even when the presupposition trigger is in a question (*Did Jordan eat some of the cake?*), as shown in Figure 1. However, in questions, classical entailments and implicatures disappear. Table 1 provides examples of triggers projecting out of several **entailment canceling operators**: negation, modals, interrogatives, and conditionals.

It is necessary to clarify in what sense presupposition is a pragmatic inference. There is no consensus on whether presuppositions should be considered part of the semantic content of expressions (see Stalnaker, 1974; Heim, 1983, for opposing views). However, presuppositions may come to be inferred via **accommodation**, a pragmatic process by which a listener infers the truth of some new fact based on its being presupposed by the speaker (Lewis, 1979). For instance, if Jordan tells Harper that *the King of Sweden wears glasses*, and Harper did not previously know that Sweden has a king, they would learn this fact by accommodation. With respect to NLI, any presupposition in the premise (short of world knowledge) will be new information, and therefore accommodation is necessary to recognize it as entailed.

3 Related Work

NLI has been framed as a commonsense reasoning task (Dagan et al., 2006; Manning, 2006). One early formulation of NLI defines "entailment" as holding for sentences p and h whenever, "typically, a human reading p would infer that h is most likely true. . . [given] common human understanding of language [and] common background knowledge" (Dagan et al., 2006). Although this sparked debate regarding the terms *inference* and *entailment*—and whether an adequate notion of "inference" could be defined (Zaenen et al., 2005; Manning, 2006; Crouch et al., 2006)—in recent work, a commonsense formulation of "inference" is widely adopted (Bowman et al., 2015; Williams et al., 2018) largely because it facilitates untrained annotators' participation in dataset creation.

NLI itself has been steadily gaining in popularity; many datasets for training and/or testing systems are now available including: FraCaS (Cooper et al., 1994), RTE (Dagan et al., 2006; Mirkin et al., 2009; Dagan et al., 2013), Sentences Involving Compositional Knowledge (Marelli et al., 2014, SICK), large scale imaging captioning as NLI (Bowman et al., 2015, SNLI), recasting other datasets into NLI (Glickman, 2006; White et al., 2017; Poliak et al., 2018), ordinal commonsense inference (Zhang et al., 2017, JOCI), Multi-Premise Entailment (Lai et al., 2017, MPE), NLI over multiple genres of written and spoken English (Williams et al., 2018, MultiNLI), adversarially filtered common sense reasoning sentences (Zellers et al., 2018, 2019, (Hella)SWAG), explainable annotations for SNLI (Camburu et al., 2018, e-SNLI), cross-lingual NLI (Conneau et al., 2018, XNLI), scientific questioning answering as NLI (Khot et al., 2018, SciTail), NLI recast-question answering (part of Wang et al. 2019, GLUE), NLI for dialog (Welleck et al., 2019), and NLI over narratives that require drawing inferences to the most plausible explanation from text (Bhagavatula et al., 2020, αNLI). Other NLI datasets are created to identify where models fail (Glockner et al., 2018; Naik et al., 2018; McCoy et al., 2019; Schmitt and Schütze, 2019), many of which are also automatically generated (Geiger et al., 2018; Yanaka et al., 2019a,b; Kim et al., 2019; Nie et al., 2019; Richardson et al., 2020).

As datasets for NLI become increasingly numerous, one might wonder, do we need yet another NLI dataset? In this case, the answer is clearly yes: despite NLI's formulation as a commonsense reasoning task, it is still unknown whether this framing has resulted in models that learn specific modes of pragmatic reasoning. IMPPRES is the first NLI dataset to explicitly probe whether models trained on commonsense reasoning actually do treat pragmatic inferences like implicatures and presuppositions as entailments without additional training on these specific inference types.

Beyond NLI, several recent works introduce resources for evaluating sentence understanding models for knowledge of pragmatic inferences. On the presupposition side, datasets such as MegaVeridicality (White and Rawlins, 2018) and CommitmentBank (de Marneffe et al., 2019) compile gradient crowdsourced judgments regarding how likely a clause embedding predicate is to trig-

ger a presupposition that its complement clause is true. White et al. (2018) and Jiang and de Marneffe (2019) find that LSTMs trained on a gradient event factuality prediction task on these respective datasets make systematic errors. Turning to implicatures, Degen (2015) introduces a dataset measuring the strength of the implicature from *some* to *not all* with crowd-sourced judgments. Schuster et al. (2020) find that an LSTM with supervision on this dataset can predict human judgments well. These resources all differ from IMPPRES in two respects: First, their empirical scopes are all somewhat narrower, as all these datasets focus on only a single class of presupposition or implicature triggers. Second, the use of gradient judgments makes it non-trivial to use these datasets to evaluate NLI models, which are trained to make categorical predictions about entailment. Both approaches have advantages, and we leave a direct comparison for future work.

Outside the topic of sentential inference, Rashkin et al. (2018) propose a new task where a model must label actor intents and reactions for particular actions described using text. Cianflone et al. (2018) create sentence-level adverbial presupposition datasets and train a binary classifier to detect contexts in which presupposition triggers (e.g., *too*, *again*) can be used.

4 Annotating MultiNLI for Pragmatics

In this section, we present results of an annotation effort that show that MultiNLI contains very little explicit evidence of pragmatic inferences of the type tested by IMPPRES. Although Williams et al. (2018) report that 22% of the MultiNLI development set sentence pairs contain lexical triggers (such as *regret* or *stopped*) in the premise and/or hypothesis, the mere presence of presupposition-triggering lexical items in the data does not show that MultiNLI contains evidence that presuppositions are entailments, since the sentential inference may focus on other types of information. To address this, we randomly selected 200 sentence pairs from the MultiNLI matched development set and presented them to three expert annotators with a combined total of 17 years of training in formal semantics and pragmatics.[4] Annotators answered the following questions for each pair: (1) are the sentences P and H related by a presupposition/implicature relation (entails/is entailed by, negated or not); (2) what subtype of inference (e.g., existence presupposition, $\langle$*some, all*$\rangle$ implicature); (3) is the presupposition trigger embedded under an entailment-cancelling operator?

Agreement among annotators was low, suggesting that few MultiNLI pairs are paradigmatic cases of implicatures or presuppositions. We found only 8 presupposition pairs and 3 implicature pairs on which two or more annotators agreed. Moreover, we found only one example illustrating a particular inference type tested in IMPPRES (the presupposition of possessed definites). All others were tagged as existence presuppositions and conversational implicatures (i.e. loose inferences dependent on world knowledge). The union of annotations was much larger: 42% of examples were identified by at least one annotator as a presupposition or implicature (51 presuppositions and 42 implicatures, with 10 sentences receiving divergent tags). However, of these, only 23 presuppositions and 19 implicatures could reliably be used to learn pragmatic inference (in 14 cases, the given tag did not match the pragmatic inference, and in 27 cases, computing the inference did not affect the relation type). Again, the large majority of implicatures were conversational, and most presuppositions were existential, and generally not linked to particular lexical triggers (e.g., topic marking).

We conclude that the MultiNLI dataset at best contains some evidence of loose pragmatic reasoning based on world knowledge and discourse structure, but almost no explicit information relevant to lexically triggered pragmatic inference, which is of the type tested in this paper.

5 Methods

Data Generation. IMPPRES consists of semi-automatically generated pairs of sentences with NLI labels illustrating key properties of implicatures and presuppositions. We generate IMPPRES using a codebase developed by Warstadt et al. (2019a) and significantly expanded for the BLiMP dataset (Warstadt et al., 2019b). The codebase, including our scripts and documentation, are publicly available.[5] Each sentence type in IMPPRES is generated according to a template that specifies the linear order of the constituents in the sentence. The constituents are sampled from a vocabulary of over 3000 lexical items annotated with grammatical features needed to ensure morphological,

[4] The full annotations are on the IMPPRES repository.

[5] github.com/alexwarstadt/data_generation

Premise	Hypothesis	Relation type	Logical label	Pragmatic label	Item type
some	*not all*	implicature (+ to −)	neutral	entailment	target
not all	*some*	implicature (− to +)	neutral	entailment	target
some	*all*	negated implicature (+)	neutral	contradiction	target
all	*some*	reverse negated implicature (+)	entailment	contradiction	target
not all	*none*	negated implicature (−)	neutral	contradiction	target
none	*not all*	reverse negated implicature (−)	entailment	contradiction	target
all	*none*	opposite	contradiction	contradiction	control
none	*all*	opposite	contradiction	contradiction	control
some	*none*	negation	contradiction	contradiction	control
none	*some*	negation	contradiction	contradiction	control
all	*not all*	negation	contradiction	contradiction	control
not all	*all*	negation	contradiction	contradiction	control

Table 2: Paradigm for the scalar implicature datasets, with ⟨*some, all*⟩ as an example.

syntactic, and semantic well-formedness. All sentences generated from a given template are structurally analogous up to the specified constituents, but may vary in sub-constituents. For instance, if the template calls for a verb phrase, the generated constituent may include a direct object or complement clause, depending on the argument structure of the sampled verb. See §6.1 and 7.1 for descriptions of the sentence types in the implicature and presupposition data.

Generating data lets us control the lexical and syntactic content so that we can guarantee that the sentence pairs in IMPPRES evaluate the desired phenomenon (see Ettinger et al., 2016, for related discussion). Furthermore, the codebase we use allows for greater lexical and syntactic variety than in many other templatic datasets (see discussion in Warstadt et al., 2019b). One limitation of this methodology is that generated sentences, while generally grammatical, often describe highly unlikely scenarios, or include low frequency combinations of lexical items (e.g., *Sabrina only reveals this pasta*). Another limitation is that generated data is of limited use for training models, since it contains simple regularities that supervised classifiers may learn to exploit. Thus, we create IMPPRES solely for the purpose of evaluating NLI models trained on standard datasets like MultiNLI.

Models. Our experiments evaluate NLI models trained on MultiNLI and built on top of three sentence encoding models: a bag of words (BOW) model, InferSent (Conneau et al., 2017), and BERT-Large (Devlin et al., 2019). The BOW and InferSent models use 300D GloVe embeddings as word representations (Pennington et al., 2014). For the BOW baseline, word embeddings for premise and hypothesis are separately summed to create sentence representations, which are concatenated to form a single sentence-pair representation which is fed to a logistic regression softmax classifier. For the InferSent model, GloVe embeddings for the words in premise and hypothesis are respectively fed into a bidirectional LSTM, after which we concatenate the representations for premise and hypothesis, their difference, and their element-wise product (Mou et al., 2016). BERT is a multilayer bidirectional transformer pretrained with the masked language modelling and next sequence prediction objectives, and finetuned on the MultiNLI dataset. We concatenate the premise and hypothesis after a special [CLS] token and separated them with the [SEP] token. The BERT representation for the [CLS] token is fed into classifier. We use Huggingface's pre-trained BERT trained on Toronto books (Zhu et al., 2015).[6]

The BOW and InferSent models have development set accuracies of 49.6% and 67.6%. The development set accuracy for BERT-Large on MultiNLI is 86.6%, similar to the results achieved by (Devlin et al., 2019), but somewhat lower than state-of-the-art (currently 90.8% on test from the ensembled RoBERTa model with long pretraining optimization, Liu et al. 2019).

6 Experiment 1: Scalar Implicatures

6.1 Scalar Implicature Datasets

The scalar implicature portion of IMPPRES includes six datasets, each isolating a different scalar implicature trigger from six types of lexical scales (of the type described in §2): determiners ⟨*some, all*⟩, connectives ⟨*or, and*⟩, modals ⟨*can, have to*⟩, numerals ⟨*2,3*⟩, ⟨*10,100*⟩, scalar adjectives, and

[6] github.com/huggingface/pytorch-pretrained-BERT/

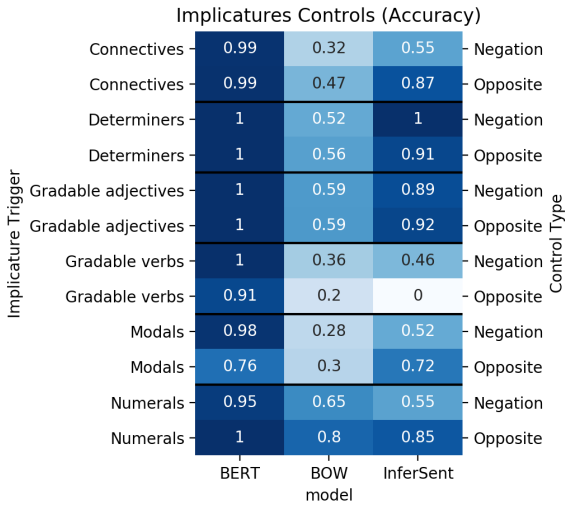

Figure 2: Results on Controls (Implicatures)

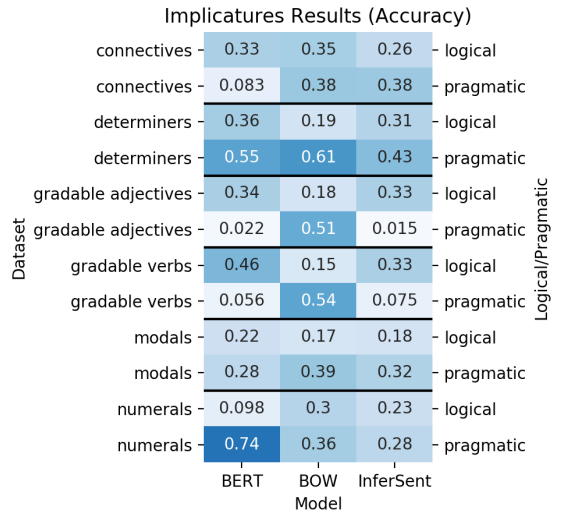

Figure 3: Results on Target Conditions (Implicatures)

verbs, e.g., ⟨*good, excellent*⟩, ⟨*run, sprint*⟩. Examples pairs of each implicature trigger can be found in Table 4 in the Appendix. For each type, we generate 100 paradigms, each consisting of 12 unique sentence pairs, as shown in Table 2.

The six target sentence pairs comprise two main relation types: 'implicature' and 'negated implicature'. Pairs tagged as 'implicature' have a premise that implicates the hypothesis (e.g., *some* and *not all*). For 'negated implicature', the premise implicates the negation of the hypothesis (e.g., *some* and *all*), or vice versa (e.g., *all* and *some*). Six control pairs are logical contradictions, representing either scalar 'opposites' (e.g., *all* and *none*), or 'negations' (e.g., *not all* and *all*; *some* and *none*), probing the models' basic grasp of negation.

As mentioned in §2.1, implicature computation is variable and dependent on the context of utterance. Thus, we anticipate two possible rational behaviors for a MultiNLI-trained model tested on an implicature: (a) be pragmatic, and compute the implicature, concluding that the premise and hypothesis are in an 'entailment' relation, (b) be logical, i.e., consider only the literal content, and not compute the implicature, concluding they are in a 'neutral' relation. Thus, we measure both possible conclusions, by tagging sentence pairs for scalar implicature with two sets of NLI labels to reflect the behavior expected under "logical" and "pragmatic" modes of inference, as shown in Table 2.

6.2 Implicatures Results & Discussion

We first evaluate model performance on the controls, shown in Figure 2. Success on these controls is a necessary condition for us to conclude that a model has learned the basic function of negation (*not*, *none*, *neither*) and the scalar relationship between terms like *some* and *all*. We find that BERT performs at ceiling on control conditions for all implicature types, in contrast with InferSent and BOW, whose performance is very variable. Since only BERT passes all controls, its results on the target items are most interpretable. Full results for all models and target conditions by implicature trigger are in Figures 8–13 in the Appendix.

For connectives, scalar adjectives and verbs, the BERT model results correspond neither to the hypothesized pragmatic nor logical behavior. In fact, for each of these subdatasets, the results are consistent with a treatment of scalemates (e.g., *and* and *or*; *good* and *excellent*) as synonyms, e.g. it evaluates the 'negated implicature' sentence pairs as 'entailment' in both directions. This reveals a coarse-grained knowledge of these meanings that lacks information about asymmetric informativity relations between scalemates. Results for modals (*can* and *have to*) are split between the three labels, not showing any predicted logical or pragmatic pattern. We conclude that BERT has insufficient knowledge of the meaning of these words.

In addition to pragmatic and logical interpretations, numerals can also be interpreted as exact cardinalities. We thus predict three different behaviors: logical "at least n", pragmatic "at least n", and "exactly n". We observe that results are inconsistent: neither the "exactly" nor "at least" interpretations hold across the board.

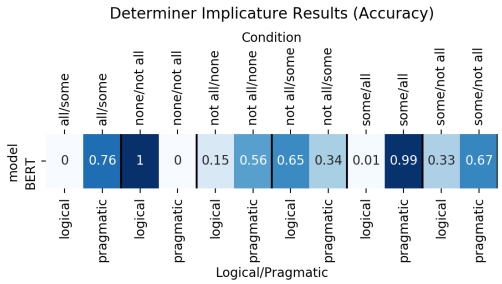

Figure 4: BERT results for scalar implicatures triggered by determiners ⟨*some, all*⟩, by target condition.

Presuppositions Premise	Hypothesis	**Label**	**Item Type**
*Trigger	Prsp	entailment	target
*Trigger	Neg. Prsp	contradiction	target
*Trigger	Neut. Prsp	neutral	target
Neg. Trigger	Trigger	contradiction	control
Modal Trigger	Trigger	neutral	control
Interrog. Trigger	Trigger	neutral	control
Cond. Trigger	Trigger	neutral	control

Table 3: Paradigm for the presupposition target (top) and control datasets (bottom). For space, *Trigger refers to either plain, Negated, Modal, Interrogative, or Conditional Triggers as per Table 1.

For the determiner dataset (*some-all*), Figure 4 breaks down the results by condition and shows that BERT behaves as though it performs pragmatic and logical reasoning in different conditions. Overall, it predicts a pragmatic relation more frequently (55% vs. 36%), and only 9% of results are consistent with neither mode of reasoning. Furthermore, the proportion of pragmatic reasoning shows consistent effects of sentence order (i.e., whether the implicature trigger is in the premise or the hypothesis), and the presence of negation in one or both sentences. Pragmatic reasoning is consistently higher when the implicature trigger is in the premise, which we can see in the results for negated implicatures: the *some–all* condition shows more pragmatic behavior compared to the *all–some* condition (a similar behavior is observed with the *not all* vs. *none* conditions).

Generally, the presence of negation lowers rates of pragmatic reasoning. First, the negated implicature conditions can be subdivided into pairs with and without negation. Among the negated ones, pragmatic reasoning is lower than for non-negated ones. Second, having negation in the premise rather than the hypothesis makes pragmatic reasoning lower: among pairs tagged as direct implicatures (*some* vs. *not all*), there is higher pragmatic reasoning with non-negated *some* in the premise than with negated *not all*. Finally, we observe that pragmatic rates are lower for *some* vs. *not all* than for *some* vs. *all*. In this final case, pragmatic reasoning could be facilitated by explicit presentation of the two items on the scale.

In sum, for the datasets besides determiners, we find evidence that BERT fails to learn even the logical relations between scalemates, ruling out the possibility of computing scalar implicatures. It remains possible that BERT could learn these logical relations with explicit supervision (see Richardson et al., 2020), but it is clear that these are not learned from training on MultiNLI. Only the determiner dataset was informative in showing the extent of the NLI BERT model's pragmatic reasoning, since it alone showed a fine-grained enough understanding of the semantic relationship of the scalemates, like *some* and *all*. In this setting BERT returned impressive results showing a high proportion of pragmatic reasoning compared to logical reasoning, which was affected by sentence order and presence of negation in a predictable way.

7 Experiment 2: Presuppositions

7.1 Presupposition Datasets

The presupposition portion of IMPPRES includes eight datasets, each isolating a different kind of presupposition trigger. The full set of triggers is shown in Table 5 in the Appendix. For each type, we generate 100 paradigms, with each paradigm consisting of 19 unique sentence pairs. (Examples of the sentence types are in Table 1).

Of the 19 sentence pairs, 15 contain target items. The first target item tests whether the model correctly determines that the presupposition trigger entails its presupposition. The next two alter the presupposition, either negating it, or replacing a constituent, leading to contradiction and neutrality, respectively. The remaining 12 show that the relation between the trigger and the (altered) presupposition is not affected by embedding the trigger under various entailment-canceling operators. 4 control items are designed to test the basic effect of entailment-canceling operators—negation, modals, interrogatives, and conditionals. In each control, the premise is a presupposition trigger embedded under an entailment-canceling operator, and the hypothesis is an unembedded sentence containing the trigger. These controls are neces-

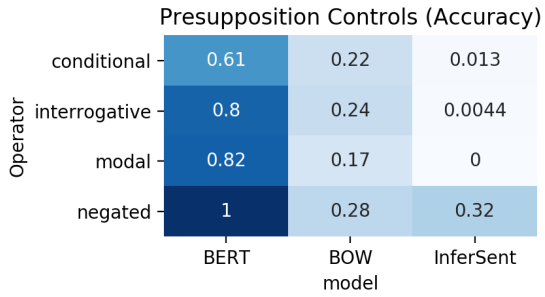

Figure 5: Results on Controls (Presuppositions).

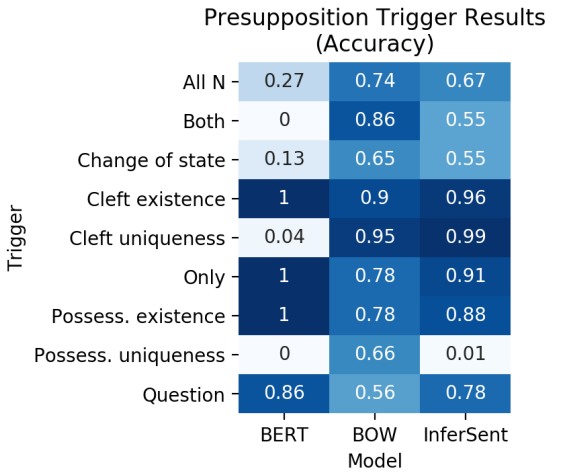

Figure 6: Results for the unembedded trigger paired with positive presupposition.

sary to establish whether models learn that presuppositions behave differently under these operators than do classical semantic entailments.

7.2 Presupposition Results & Discussion

The results from presupposition controls are in Figure 5. BERT performs well above chance on each control (acc. $>$ 0.33), whereas BOW and InferSent perform at or below chance. In the "negated" condition, BERT correctly identifies that the trigger is contradicted by its negation 100% of the time, e.g., *Jo's cat didn't go* contradicts *Jo's cat went*. In the other conditions, it correctly identifies the neutral relation the majority of the time, e.g., *Did Jo's cat go?* is neutral with respect to *Jo's cat went*. This indicates that BERT mostly learns that negation, modals, interrogatives, and conditionals cancel classical entailments, while BOW and InferSent do not capture the ordinary behavior of these common operators.

Next, we test whether models identify presuppositions of the premise as entailments, e.g., that *Jo's cat went* entails that *Jo has a cat*. Recall from §2.2 that this is akin to a listener accommodating a presupposition. The results in Figure 6 show that each of the three models accommodates some presuppositions, but this depends on both the nature of the presupposition and the model. For instance, the BOW and InferSent models accommodate presuppositions of nearly all trigger types at well above chance rates (acc. $\gg$ 33%). For the uniqueness presupposition of clefts, these models generally correctly predict an entailment (acc. $>$ 90%), but for most triggers, performance is less reliable. By contrast, BERT's behavior is bimodal. It always accommodates the existence presuppositions of clefts and possessed definites, as well as the presupposition of *only*, but almost never accommodates any presupposition involving numeracy, e.g. *Both flowers that bloomed died* entails *There are exactly two flowers that bloomed.*[7]

Finally, we evaluate whether models predict that presuppositions project out of entailment canceling operators (e.g., that *Did Jo's cat go?* entails that *Jo has a cat*). We can only consider such a prediction as evidence of projection if two conditions hold: (a) the model correctly identifies that the relevant operator cancels entailments in the control from the same paradigm (e.g., *Did Jo's cat go?* is neutral with respect to *Jo's cat went*), and (b) the model identifies the presupposition as an entailment when the trigger is unembedded in the same paradigm (e.g. *Jo's cat went* entails *Jo has a cat*). Otherwise, a model might correctly predict entailment essentially by accident if, for instance, it systematically ignores negation. For this reason, we filter out results for the target conditions that do not meet these criteria.

Figure 7 shows results for the target conditions after filtering. While InferSent rarely predicts that presuppositions project, we find strong evidence that the BERT and BOW models do. Specifically, they correctly identify that the premise entails the presupposition (acc. $\geq$ 80% for BERT, acc. $\geq$ 90% for BOW). Furthermore, BERT is the only model to reliably identify (i.e., over 90% of the time) that the negation of the presupposition is contradicted. These results hold irrespective of the entailment canceling operator. No model reliably performs above chance when the presupposition is altered to be neutral (e.g., *Did Jo's cat go?* is neu-

[7]The presence of *exactly* might contribute to poor performance on numeracy examples. We suspect MultiNLI annotators may have used it disproportionately for neut. hypotheses.

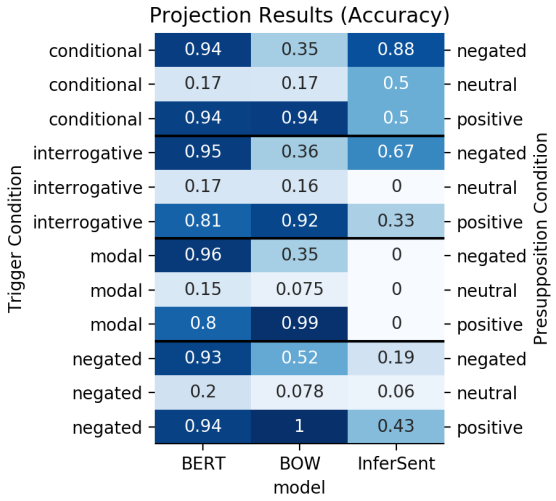

Figure 7: Results for presupposition target conditions involving projection.

tral with respect to *Jo has a cat*).

It is surprising that the simple BOW model can learn some of the projective behavior of presuppositions. One explanation for this finding is that many of the key features of presupposition projection are insensitive to word order. If a lexical presupposition trigger is present at all in a sentence, a presupposition will generally arise irrespective of its position in the sentence. There are some edge cases where this heuristic is insufficient, but IMPPRES is not designed to test such cases.

To summarize, training on NLI is sufficient for all models we evaluate to learn to accommodate presuppositions of a wide variety of unembedded triggers, though BERT rejects presuppositions involving numeracy. Furthermore, BERT and even the BOW model appear to learn the characteristic projective behavior of some presuppositions.

8 General Discussion & Conclusion

We observe some encouraging results in §6–7. We find strong evidence that BERT learns scalar implicatures associated with determiners *some* and *all*. Pragmatic or logical reasoning was not diagnosable for the other scales, whose meaning was not fully understood by our models (as most scalar pairs were treated as synonymous). In the case of presuppositions, the BERT NLI models, and BOW to some extent, perform well on a number of our subdatasets (*only*, cleft existence, possessive existence, questions). For the other subdatasets, the models did not perform as expected on the basic unembedded presupposition triggers, again suggesting the model's lack of knowledge of the basic meaning of these words. Though their behavior is far from systematic, this is suggestive evidence that some NLI models can perform in ways that correlate with human-like pragmatic behavior.

Given that MultiNLI contains few examples of the type found in IMPPRES (see §4), where might our positive results come from? There are two potential sources of signal for the BERT model: NLI training, and pretraining (either BERT's masked language modeling objective or its input word embeddings). NLI training provides specific examples of valid (or invalid) inferences constituting an incomplete characterization of what commonsense inference is in general. Since presuppositions and scalar implicatures triggered by specific lexical items are largely absent from the MultiNLI data used for NLI training, any positive results on IMPPRES would likely use prior knowledge from the pretraining stage to make an inductive leap that pragmatic inferences are valid commonsense inferences. The natural language text used for pretraining certainly contains pragmatic information, since, like any natural language data, it is produced with the assumption that readers are capable of pragmatic reasoning. Maybe this induces patterns in the data that make the nature of those assumptions recoverable from the data itself.

This work is an initial step towards rigorously investigating the extent to which NLI models learn semantic versus pragmatic inference types. We have introduced a new dataset IMPPRES for probing this question, which can be reused to evaluate pragmatic performance of any NLI given model.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 1850208 awarded to A. Warstadt. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF. Thanks to the FAIR NLP & Conversational AI Group, the Google AI NLP group, and the NYU ML[2], including Sam Bowman, He He, Phu Mon Htut, Katharina Kann, Haokun Liu, Ethen Perez, Richard Pang, Clara Vania for discussions on the topic, and/or feedback on an earlier draft. Additional thanks to Marco Baroni, Hagen Blix, Emmanuel Chemla, Aaron Steven White, and Luke Zettlemoyer for insightful comments.

References

Aristotle. *De Interpretatione*.

David I. Beaver. 1997. Presupposition. In *Handbook of logic and language*, pages 939–1008. Elsevier.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *Proceedings of the 2020 International Conference on Learning Representations (ICLR)*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *Advances in Neural Information Processing Systems*, pages 9539–9549.

Andre Cianflone, Yulan Feng, Jad Kabbara, and Jackie Chi Kit Cheung. 2018. Let's do it "again": A first computational approach to detecting adverbial presupposition triggers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2747–2755, Melbourne, Australia. Association for Computational Linguistics.

Herbert H Clark. 1996. *Using language*. Cambridge University Press.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485. Association for Computational Linguistics.

Robin Cooper, Richard Crouch, Jan Van Eijck, Chris Fox, Josef Van Genabith, Jan Jaspers, Hans Kamp, Manfred Pinkal, Massimo Poesio, Stephen Pulman, et al. 1994. FraCaS: A framework for computational semantics. *Deliverable D6*.

Richard Crouch, Lauri Karttunen, and Annie Zaenen. 2006. Circumscribing is not excluding: A reply to Manning. Ms., Palo Alto Research Center.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

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

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Gerald Gazdar. 1979. *Pragmatics, implicature, presupposition and logical form*. Academic Press, NY.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2018. Stress-testing neural models of natural language inference with multiply-quantified sentences. In *CoRR*.

Oren Glickman. 2006. *Applied textual entailment*. Bar-Ilan University.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655. Association for Computational Linguistics.

H Paul Grice. 1975. Logic and conversation. *1975*, pages 41–58.

Irene Heim. 1983. On the projection problem for presuppositions. *Formal semantics: The essential readings*, pages 249–260.

Laurence Horn. 1989. *A natural history of negation*. University of Chicago Press.

Laurence R Horn and Gregory L Ward. 2004. *The handbook of pragmatics*. Wiley Online Library.

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*, pages 4208–4213, Florence, Italy. Association for Computational Linguistics.

Lauri Karttunen. 1973. Presuppositions of compound sentences. *Linguistic inquiry*, 4(2):169–193.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *Proceedings of Association for the Advancement of Artificial Intelligence (AAAI)*.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Alice Lai, Yonatan Bisk, and Julia Hockenmaier. 2017. Natural language inference from multiple premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 100–109, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

David Lewis. 1979. Scorekeeping in a language game. In *Semantics from different points of view*, pages 172–187. Springer.

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*.

Christopher D. Manning. 2006. Local textual inference: It's hard to circumscribe, but you know it when you see it – and NLP needs it. Ms., Stanford University.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. In *Proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Shachar Mirkin, Roy Bar-Haim, Ido Dagan, Eyal Shnarch, Asher Stern, Idan Szpektor, and Jonathan Berant. 2009. Addressing discourse and document structure in the RTE search task. In *Textual Analysis Conference*.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 130–136, Berlin, Germany. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019. Analyzing compositionality-sensitivity of NLI models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6867–6874.

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, Doha, Qatar. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

Christopher Potts. 2015. Presupposition and implicature. *The handbook of contemporary semantic theory*, 2:168–202.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind:

Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Kyle Richardson, Hai Hu, Lawrence S Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*.

Bertrand Russell. 1905. On denoting. *Mind*, 14(56):479–493.

Martin Schmitt and Hinrich Schütze. 2019. SherLIiC: A typed event-focused lexical inference benchmark for evaluating natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 902–914, Florence, Italy. Association for Computational Linguistics.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the richness of the linguistic signal in predicting pragmatic inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Robert Stalnaker. 1974. Pragmatic presuppositions. In Milton K. Munitz and Peter K. Unger, editors, *Semantics and Philosophy*, pages 135–148. New York University Press.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Interantional Conference on Learning Representations (ICLR)*.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretič, and Samuel R. Bowman. 2019a. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of EMNLP-IJCNLP*, pages 2870–2880.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2019b. BLiMP: The benchmark of linguistic minimal pairs for English. *arXiv preprint arXiv:1912.00582*.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

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, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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, 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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.

Annie Zaenen, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: Can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 31–36, Ann Arbor, Michigan. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Appendix

Type	Premise	Hypothesis
Connectives	These cats or those fish appear.	These cats and those fish don't both appear.
Determiners	Some skateboards tipped over.	Not all skateboards tipped over.
Numerals	Ten bananas were scorching.	One hundred bananas weren't scorching.
Modals	Jerry could wake up.	Jerry didn't need to wake up.
Scalar adjectives	Banks are fine.	Banks are not great.
Scalar verbs	Dawn went towards the hills.	Dawn did not get to the hills.

Table 4: The scalar implicature triggers in IMPPRES. Examples are automatically generated sentences pairs from each of the six datasets for the scalar implicatures experiment. The pairs belong to the "Implicature ($+$ to $-$)" condition.

Type	Premise (Trigger)	Hypothesis (Presupposition)
All N	All six roses that bloomed died.	Exactly six roses bloomed.
Both	Both flowers that bloomed died.	Exactly two flowers bloomed.
Change of State	The cat escaped.	The cat used to be captive.
Cleft Existence	It is Sandra who disliked Veronica.	Someone disliked Veronica.
Cleft Uniqueness	It is Sandra who disliked Veronica.	Exactly one person disliked Veronica.
Only	Only Lucille went to Spain.	Lucille went to Spain.
Possessed Definites	Bill's handyman won.	Bill has a handyman.
Question	Sue learned why Candice testified.	Candice testified.

Table 5: The presupposition triggers in IMPPRES. Examples are automatically generated sentences pairs from each of the eight datasets for the presupposition experiment. The pairs belong to the "Plain Trigger / Presupposition" condition.

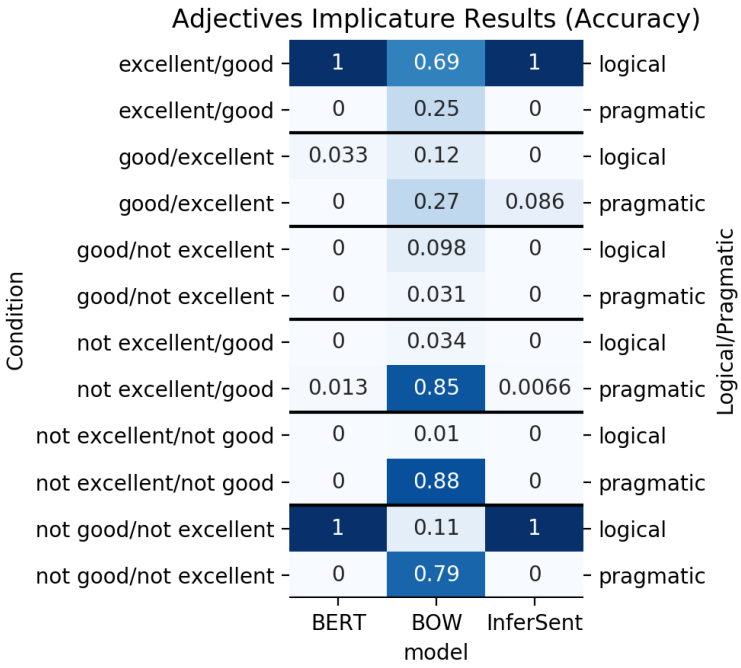

Figure 8: Results for the scalar implicatures triggered by adjectives, by target condition.

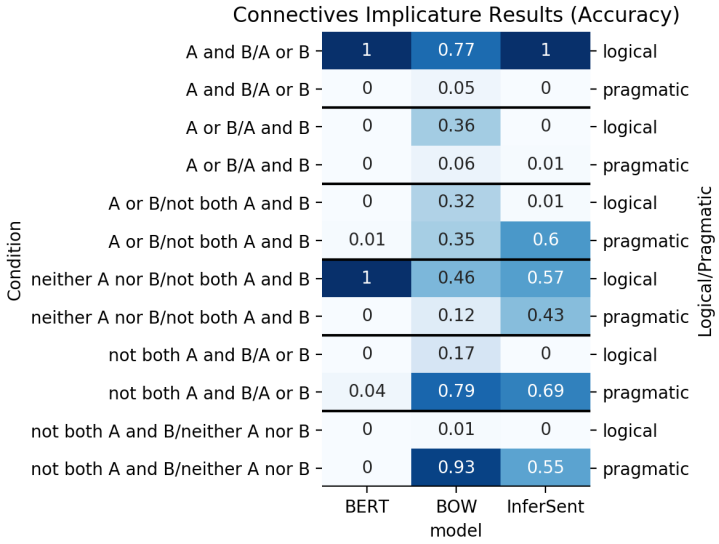

Figure 9: Results for the scalar implicatures triggered by adjectives, by target condition.

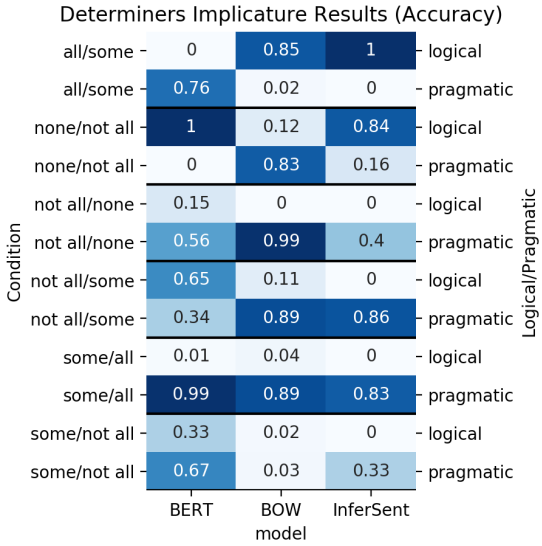

Figure 10: Results for the scalar implicatures triggered by determiners, by target condition.

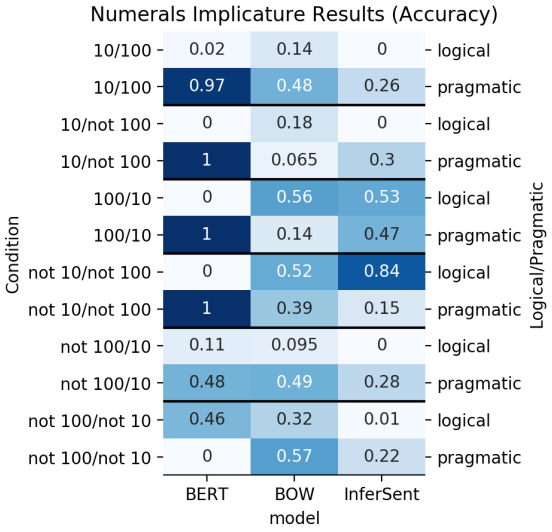

Figure 12: Results for the scalar triggered by numerals, by target condition.

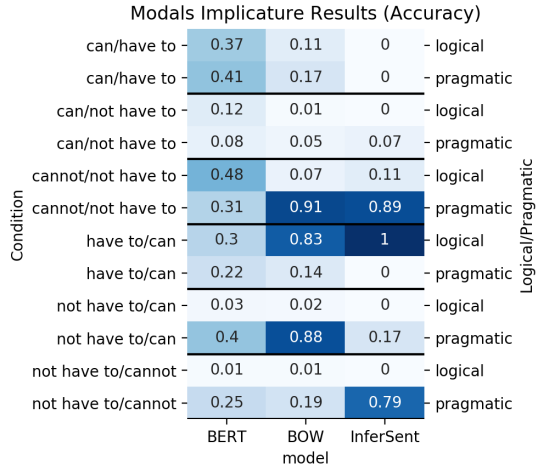

Figure 11: Results for the scalar implicatures triggered by modals, by target condition.

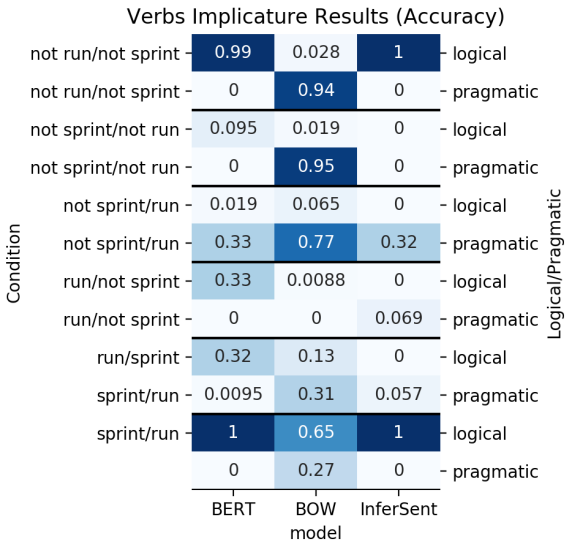

Figure 13: Results for the scalar implicatures triggered by verbs, by target condition.