Evaluating the Projectivity of Presupposition Triggers in Various Entailment-Canceling Environments

Anonymous ACL submission

Abstract

Previous studies investigate the ability of models to make pragmatic inferences using presupposition triggers. However, although projection of presuppositions can vary depending on the combination of triggers and environments, they evaluate the performance of models without human baseline, or include only negative sentences as entailment-canceling environments. To evaluate inferences with presupposition triggers, it is necessary to solicit human judgments as a baseline for model evaluation and use various types of entailment-canceling environments. In this study, we introduce a template-based natural language inference dataset called Projectivity of Presupposition Triggers (PPT), which includes 9,800 sentence pairs crossed with six types of presupposition triggers and four types of syntactic environments. Analysing judgements from 283 people on a subset of the dataset, we find that humans take most presupposition patterns as projective, but the projectivity varies depending on the combination of triggers and environments. In contrast, models judge some patterns as non-projective, indicating that the ability of the models to process presuppositions may not be human-like. This result highlights that researchers working on model evaluation and dataset creation need to take extra care of the combination of presupposition triggers and environments where they are embedded.

1 Introduction

There is an open question whether linguistic models can make pragmatic inferences in the same way that humans do (Pavlick, 2022). In particular, it remains open whether machines can make one type of pragmatic inference, presupposition. Presupposition refers to a relation between two sentences (Stalnaker, 1974; Beaver, 1997). It is often triggered by linguistic expressions called presupposition triggers such as *again*, as shown in Figure 1. Although presupposition appears similar to another

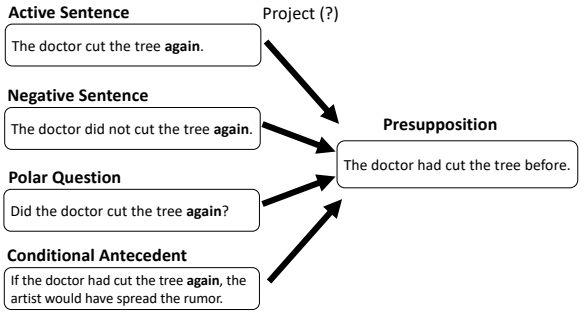

Figure 1: Projectivity of presupposition. It is assumed that presuppositions of triggers project out of entailment-canceling environments. However, projectivity of presuppositions can vary, depending on the triggers and syntactic environments where they are embedded.

relation, entailment, these two inferences are distinct. Presupposition is assumed to project out of *entailment-canceling environments* (e.g., a negative sentence, question, and conditional antecedent) whereas entailment cannot.

However, the presupposition does not always project. The projectivity of a presupposition varies depending on factors such as context, a lexical item, prior beliefs, and the social identify of the speaker (Karttunen, 1971; Simons, 2001; Tonhauser et al., 2018; Degen and Tonhauser, 2021b). Considering that most previous studies focus on clause-embedding predicates and presuppositions of other triggers are assumed to project, variable projective behaviors of presuppositions are understudied. For instance, presupposition triggered by *again* can more likely project out of a question than a negative sentence.

Previous studies in natural language processing examine the performance of models on presuppositions by using a natural language inference (NLI) task (Ross and Pavlick, 2019; Jeretic et al., 2020; Parrish et al., 2021). In the NLI task, one determines whether a premise sentence entails or contradicts a hypothesis sentence (Dagan et al., 2006; Bowman et al., 2015). However, each study

has some limitations. For instance, Jeretic et al. (2020) do not conduct human evaluation as a baseline. de Marneffe et al. (2019) shows that human judgments on presupposition of clause-embedding predicates vary. Given this, it is likely that projectivity of presuppositions also varies in other trigger cases. Therefore, when assessing the performance of models on presuppositions, it is necessary to solicit human intuitive judgments to make a baseline. Parrish et al. (2021) use only negative sentences for entailment-canceling environments; hence, it remains unclear whether models take presuppositions as projective out of any entailment-canceling environments.

In order to address these concerns, in this work, we use various entailment-canceling environments to investigate projectivity of presupposition and test whether human judgments on presupposition projection are variable depending upon the combination of triggers and environments where they appear. In addition, we use the human result for a baseline to evaluate models' performance in detail.

To this end, we introduce a new evaluation dataset, Projectivity of Presupposition Triggers (PPT), which consists of 9,800 sentence pairs that are generated with templates and is semi-automatically generated with templates and is designed to test the performance of NLI models on sentences with presupposition triggers. Our dataset includes six trigger types crossed with four syntactic environments, making it possible to investigate the performance of models for a wide range of sentence patterns. Furthermore, we analyze human intuitive judgments on a subset of this dataset (480 sentence pairs) from 283 people (56.6 people on average per sentence pair) to examine the variable projective behaviors of presuppositions, and evaluate four models (Bag-of-Words (BOW), InferSent (Conneau et al., 2017), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2020)) against the human performance.

By analyzing the human judgment data, we find that humans take most presuppositions as projective with some variability, but transformer-based models judge some patterns as non-projective. With this finding, we conclude that the ability of models to process presuppositions is not human-like yet and researchers working on model evaluation and dataset creation need to take extra care of the combination of various triggers and syntactic environments to investigate the ability of models to process presuppositions.

In conclusion, this study makes the following contributions:[1]

- We introduce a novel evaluation dataset PPT to test the capability of the model for processing presuppositions of different triggers embedded under various entailment-canceling environments.
- Through our human intuitive judgment experiment, we find that the projectivity of presupposition depends on the combination of presupposition triggers and entailment-canceling environments where they are embedded.
- We demonstrate that models are incapable of making sophisticated human-like pragmatic inferences for presupposition triggers.

2 Background

2.1 Presupposition in Pragmatics

This study focuses on one type of pragmatic inference: presupposition. Presupposition is a pragmatic relation between two sentences and is considered to be taken granted by speakers (Stalnaker, 1974; Beaver, 1997). Presuppositions are often triggered by lexical items called presupposition triggers. Figure 1 illustrates *again* as a presupposition trigger. There are various types of presupposition triggers such as manner adverbs, comparatives, and temporal adverbs (see Levinson (1983), Beaver (1997), and Potts (2015) for a list of presupposition triggers).

A property that makes presupposition distinct from other inter-sentential relations is projection: presupposition survives in environments such as questions and negation (Karttunen, 1973; Heim, 1983). For instance, the presupposition of the sentence *the doctor cut the tree again last night* is assumed to hold in its question (*did the doctor cut the tree last night?*) and negation (*the doctor did not cut the tree again last night*) counterparts. In contrast, in these environments, entailment for the same sentence *the doctor cut the tree last night* disappears. The environments such as questions and negation are called entailment-canceling environments. Models can process presupposition triggers only if they correctly infer presuppositions

[1] We will make our dataset and codebase publicly available.

Trigger Type	Example Triggers	Example Premise
Iterative	*again*	The assistant split the log **again**.
Change-of-state verb	*stop*, *quit*, *finish*	The assistant **stopped** splitting the log.
Manner adverb	*quietly*, *slowly*, *angrily*	The assistant split the log **quietly**.
Factive verb	*remember*, *regret*, *forget*	The assistant **remembered** splitting the log.
Comparative	*better*, *earlier*, *more seriously*	The assistant split the log **better than** the girl.
Temporal adverb	*before*, *after*, *while*	The assistant split the log **before** bursting into the room.

Table 1: Examples of presupposition triggers with an active premise.

to project out of the entailment-canceling environments.

However, the projection of presupposition is not straightforward. The projectivity of presupposition can vary depending on factors such as context, a lexical item, prior beliefs, and the social identify of the speaker (Karttunen, 1971; Simons, 2001; Tonhauser et al., 2018; Degen and Tonhauser, 2021b). However, most previous research exclusively focuses on factive predicates (e.g., *know*, *remember*), making it unclear whether other triggers show the same variability of projection. For instance, does the presupposition of *change-of-state verbs* project out of negation in the same manner as the presupposition of *again*? Therefore, this study investigates whether different types of presupposition triggers show variable projectivity depending on entailment-canceling environments where they appear by soliciting human judgments. We then evaluate models against the human result to examine whether models show the same type of variability as humans.

2.2 Presuppositions in NLI

Some NLI datasets are introduced to evaluate the ability of models to make pragmatic inferences (Ross and Pavlick, 2019; Jeretic et al., 2020; Parrish et al., 2021). human judgments.

IMPPRES (Jeretic et al., 2020) is a template-based dataset designed to test two types of pragmatic inferences: implicature and presupposition. Using this dataset, Jeretic et al. (2020) find that although models (e.g., BERT (Devlin et al., 2019)) fail to make pragmatic inferences in some cases, they learn the projective behavior of some presuppositions. However, Jeretic et al. (2020)'s conclusion is not persuasive, considering that they do not conduct human evaluation on the dataset. Humans are known to often make seemingly unsystematic judgments about projection on both natural (Ross and Pavlick, 2019; de Marneffe et al., 2019) and controlled (White and Rawlins, 2018) items, which makes it difficult to interpret the model performance by any explicit definition rather than human judgment results. Following Parrish et al. (2021), this study conducts a human judgment experiment to obtain a baseline for model evaluation.

NOPE (Parrish et al., 2021) includes naturally-occurring data with presupposition triggers. With this dataset, Parrish et al. (2021) evaluate BERT-based models against human performance, finding that the models process presupposition triggers in the same way as humans even when they are embedded under negation. However, one limitation of NOPE is that it includes only one entailment-canceling environment, negation. To make a stronger conclusion about the model performance on projection of presuppositions, we need to include more types of entailment-canceling environments besides negation. Following Jeretic et al. (2020), the PPT dataset includes not only negation but also a polar question and conditional antecedent as entailment-canceling environments.

3 Data Generation

3.1 Presupposition Triggers and Syntactic Environments

We use six types of presupposition triggers: 1) an iterative *again*, 2) change-of-state verbs (CSV), 3) manner adverbs, 4) factive verbs, 5) comparatives, and 6) temporal adverbs, as shown in Table 1. We select these triggers from the lists made by Levinson (1983) and Potts (2015), because they are not included in the previous template-based dataset IMPPRES (Jeretic et al., 2020) and can be easily incorporated into templates.

For syntactic environments where presupposition triggers occur, we use four environments: 1) an active sentence, 2) a negative sentence, 3) a polar question, and 4) an antecedent of a counterfactual conditional, as exemplified in Table 2. Unlike Jeretic et al. (2020), we do not include a modal environment in our dataset to prevent explosion of

Construction	Premise	Hypothesis	Label
Active sentence	The doctor cut the tree **again**.	The doctor had (not) cut the tree before.	E (C)
Negative sentence	The doctor did not cut the tree **again**.	The doctor had (not) cut the tree before.	E (C)
Polar question	Did the doctor cut the tree **again**?	The doctor had (not) cut the tree before.	E(C)
Counterfactual conditional	If the doctor had cut the tree **again**, the dancer could have burst into the room.	The doctor had (not) cut the tree before.	E (C)

Table 2: Examples of premise-hypothesis pairs with *again*. *E* and *C* stand for *Entailment* and *Contradiction*, respectively. These labels are assumed to be assigned if presuppositions project. Eight conditions are generated for each of eight triggers (48 conditions in total).

Construction	Trigger	Templates	Examples
Counterfactual conditional	*Manner adverb*	*P*: If the N_1 had VP_1 MAdv, the N_2 Modal have VP_2. H_1: The N_1 VP_1. H_2: The N_1 did not VP_1.	*P*: If the girl had set the dish on the table slowly, the boy could have burst into the room. H_1: The girl set the dish on the table. H_2: The girl did not set the dish on the table.

Table 3: Examples of templates and sentences with a manner adverb in PPT.

the dataset. The active sentence is used as a control condition to test whether models can process presupposition triggers in the simple case. The other three conditions are entailment-canceling environments, which serve as target conditions. These are used to test variable projectivity of presupposition of triggers. Each trigger type occurs in four environments, thus constituting eight premise-hypothesis pairs each in total. For instance, Table 2 shows examples of premise-hypothesis pairs with *again*. We generate 100 sentences for each premise-hypothesis pair pattern, respectively. Therefore, the PPT comprises 4,800 target sentence pairs.

We make a control condition for each condition where a hypothesis is the affirmative or negative version of its premise. For instance, in the control condition of the negation with *again*, if the premise is *the doctor did not cut the tree again*, its hypothesis is *the doctor cut the tree again* or *the doctor did not cut the tree again*. These statements would be labeled as contradiction and entailment, respectively. These control sentences are used as sanity check in human evaluation and to ensure that the model performance is not affected by the mere presence of presupposition triggers. We generate 100 sentence pairs for each control condition; therefore, the PPT includes 4,800 control sentences. In total, PPT comprises 9,600 sentence pairs.

3.2 Templates

Sentence pairs in PPT are automatically generated with templates on the basis of the codebase developed by Yanaka and Mineshima (2021). We use template-based data instead of naturally-occurring data for our dataset to control the plausibility of the sentences. Previous work (Karttunen, 1971; Simons, 2001; Tonhauser et al., 2018) shows that the projectivity of presupposition varies depending on its content. For instance, the sentence *John didn't stop going to the restaurant* leads to the inference that John had been going to the restaurant before. In contrast, the sentence *John didn't stop going to the moon* is less likely to yield the inference that John had been going to the moon before. This difference can be attributed to our world knowledge: it is more plausible for one to go to the restaurant than go to the moon in the real world. Because the effect of plausibility of the content is not within the scope of this study, we use templates to control it.

Examples of the templates and sentences are given in Table 3. An index is assigned to N and VP to distinguish their multiple occurrences, if required. For a verb in VP, we use verbs having the same form in past tense and past participle forms (e.g., set, cut, burst) to ensure that the morphological difference between premise and hypothesis sentences is as small as possible. [2]

4 Human Evaluation

To investigate the extent to which the projection behavior of presupposition varies depending on the combination of presupposition triggers and syn-

[2] A full list of the templates and their example sentences is provided in Appendix A.

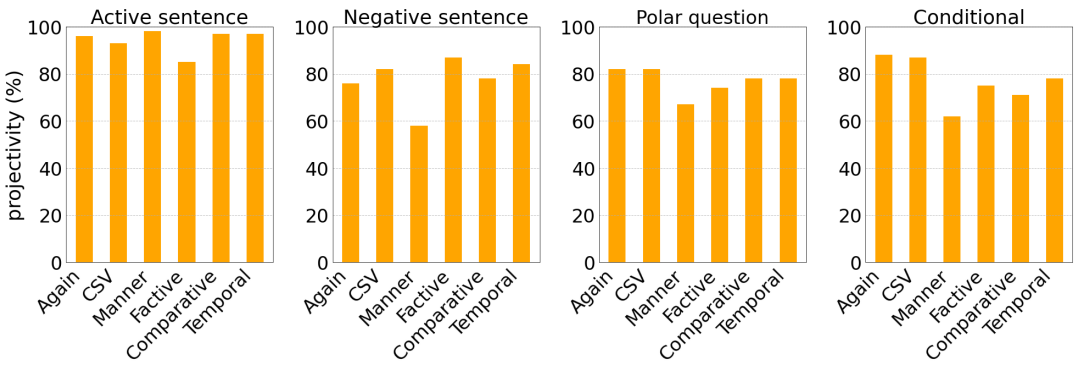

Figure 2: Results on target conditions for humans.

tactic environments where they are embedded, we conduct a human judgment experiment.

4.1 Design and Procedure

We randomly select 10 out of 100 sentence pairs from each target condition and select 2 sentence pairs from each control condition, extracting 480 and 96 items, respectively.

Using Amazon Mechanical Turk,[3] we recruit 635 people with the requirements of having an approval rating of 99% or higher, having at least 5,000 approved tasks, and being located in the US, the UK, or Canada. Among them, we exclude the responses of 352 participants from our data analysis because their accuracy on control conditions is below or equal to 90%. We analyze the judgment data of the remaining 283 participants. As a result, the items are judged by 56.6 participants on average. We make sure that the workers are paid at least $12.0 USD per hour. More information about our experiment is reported in Appendix B.

For the results on target conditions, we use the term, projectivity, rather than accuracy. As suggested by previous research on clause-embedding predicates (Tonhauser et al., 2018; Degen and Tonhauser, 2021a,b), human intuitive judgments on projectivity can vary. Therefore, we cannot assume any predetermined accurate label for sentence pairs. Projectivity is calculated based on whether presupposition projects. For instance, if a participant labels the hypothesis *the singer cut the tree* as entailment given the premise *the singer did not cut the tree slowly*, the response is considered projective. Taking another example, if the hypothesis *the boy did not burst into the room* is judged as contradiction given the premise *the boy did not burst into the room more seriously than the singer*, it is considered projective. Otherwise, these two examples are considered non-projective.

[3] https://www.mturk.com

Condition	Accuracy (%)
Active sentence	99.9
Negative sentence	99.3
Polar question	51.2
Counterfactual conditional	93.1

Table 4: Human performance on control conditions.

4.2 Results and Discussion

Table 4 shows human accuracy on the control conditions. Accuracy on the active sentence, negative sentence, and counterfactual conditional is at the ceiling (99.9%, 99.3%, and 93.1%, respectively). In contrast, the performance on the polar question condition is better than that of random choice (51.2% over a 33.3%) but poorer than the other three conditions. This might be because it is hard to imagine the situation based on a polar question (e.g., *did the singer put the book on the shelf more seriously than the boy?*), since the primary function of the polar question is to ask the truth of its content. The distribution of the other responses to the question condition is as follows: 38.4% and 10.2% for entailment and contradiction, respectively.

Figure 2 shows the results on the target conditions. The performance on the active sentence shows that presupposition holds in active sentences. One exception is the factive verb condition whose projectivity is less than 90% unlike other triggers (84.6%). This result is consistent with the previous findings that presupposition of factive predicates varies (Tonhauser et al., 2018; Degen and Tonhauser, 2021a,b).

In the other three entailment-canceling environ-

ment conditions, projectivity exceeds the chance level (33.3%), which indicates that presuppositions of triggers survive in these three conditions, too. As shown in Figure 2, the projectivity varies depending on triggers and environments. For instance, presuppositions of manner adverbs are less likely to project out of the negative sentence (e.g., *The kid did not bet $100 on the race quickly.*) (58.3%), polar question (e.g., *Did the stranger burst into the room angrily?*) (66.6%), and counterfactual conditional (e.g., *If the woman had slit the envelope easily, the director would have upset the boat.*) (62.0%) than the active sentence (e.g., *The boy read the letter quietly.*) (98.2%). In addition, presuppositions of the comparative are less likely to project out of the negative sentence (e.g., *The worker did not thrust the fork into the cake better than the girl.*) (64.1%), question (e.g., *Did the director cast bronze into a statue anxiously?*) (77.9%), and conditional (e.g., *If the doctor had let the blinds down earlier than the student, the stranger could have shut the door.*) (70.7%) than the active sentence (e.g., *The child spread the rumor better than the doctor.*) (97.0%). These variable projection behaviors suggest that the projectivity of presupposition depends on the combination of presupposition triggers and syntactic environments where they are embedded. Previous work (Degen and Tonhauser, 2021a) shows that the projectivity of presupposition of clause-embedding predicates varies. The results suggest that the similar types of variability also exists among different types of presupposition triggers.

In order to investigate whether the variable projection behaviors are attributed to each item within each condition, we look at the standard deviation of the projectivity in each condition. The mean standard deviations are 12.9, 7.2, and 10.3, for the manner adverb embedded under the negative sentence, polar question, and counterfactual conditionals, respectively. They are 10.7, 7.1, and 10.5, for the comparative embedded under the negative sentence, polar question, and counterfactual conditionals, respectively. Given these standard deviations, the projectivity of presuppositions in these conditions are variable within each trigger condition. Some items have different projectivity despite the fact that they use the same trigger (e.g., 48.3% and 70.4% for *the dancer did not hit the ball with the bat earlier than the girl* and *the teacher did not split the log earlier than the dancer*, respectively). This result indicates that not only each trigger item but also other factors such as other lexical items in the sentence and the plausibility of the sentence affect projectivity. We leave it for the future work what factor other than individual triggers affects projectivity.

In summary, a detailed analysis reveals variable projection behaviors in some conditions, indicating that projectivity of presupposition depends on the combination of triggers and environments.

5 Model Evaluation

We evaluate standard NLI models against the PPT dataset. To investigate whether the model performance on presuppositions mirrors human performance, we compare the model results with those of humans.

5.1 Models

We evaluate four models: a bag-of-words (BOW) model, an InferSent model (Conneau et al., 2017), RoBERTa-base (Liu et al., 2019), and DeBERTa-large (He et al., 2020). For the first two models, we follow Parrish et al. (2021)'s implementation[4] and use MNLI (Williams et al., 2018) to fine-tune the parameters. We use the GloVe embeddings for the word-level representations (Pennington et al., 2014). For the remaining two transformer-based models, we use Huggingface's (Wolf et al., 2020) pretrained RoBERTa-base and DeBERTa-large fine-tuned on MNLI.

5.2 Results and Discussion

Figure 3 shows the model's performance on control conditions along with human performance. InferSent and BOW models perform poorly on the control conditions, compared with two-transformer models and humans. Their accuracy does not reach 75% (ranges 6.3–67.9% and 0.0–74.9% for BOW and InferSent, respectively) RoBERTa and DeBERTa, respectively). Similar to humans, RoBERTa and DeBERTa achieve performance at the ceiling, indicating that the mere presence of presupposition triggers in various syntactic environments does not affect their performance. However, one exception is the polar question, where RoBERTa performs at the chance level (31.8%); similar to humans, DeBERTa shows moderate accuracy (50.0%). This indicates that assigning the

[4] https://github.com/nyu-mll/nope

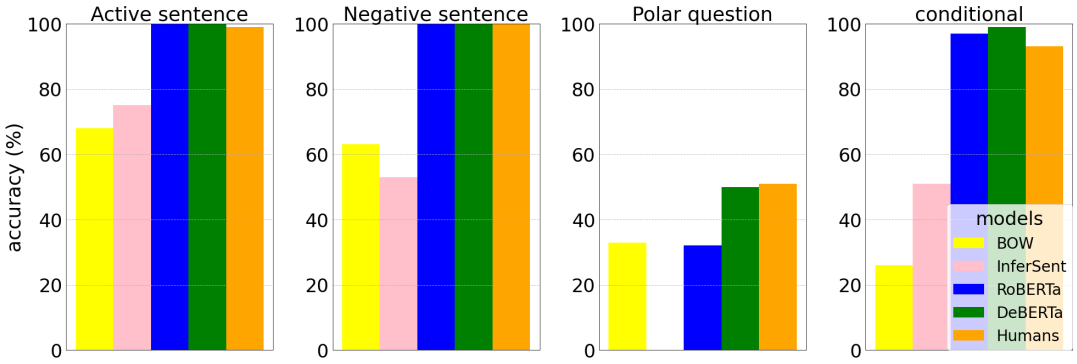

Figure 3: Results on control conditions for four models and humans.

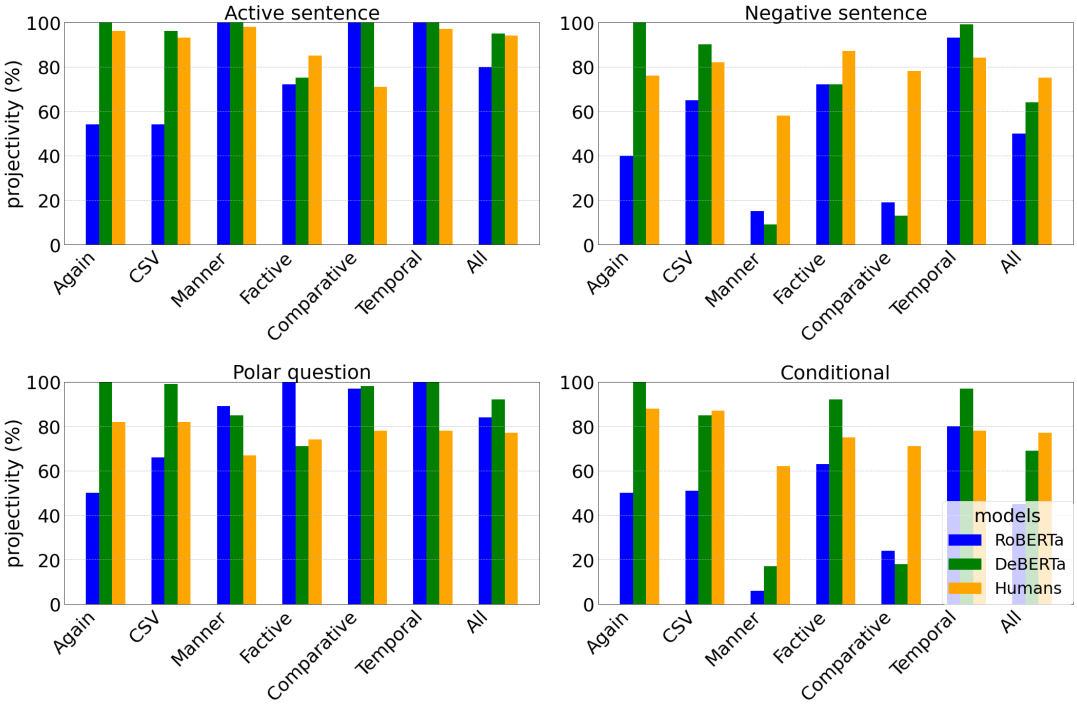

Figure 4: Results on target conditions for two transformer-based models and humans.

neutral label to the polar question condition is difficult for transformer-based models and humans. This difficulty might be attributed to the task: it is hard to imagine a situation where a polar question is true because its primary function is to ask the truth of its content.

Figure 4 shows results of RoBERTa and DeBERTa for target conditions with those of humans. Similar to humans, RoBERTa and DeBERTa take presuppositions of most triggers as projective with some exceptions as follows.

Unlike humans and DeBERTa, RoBERTa takes *again* and CSV as moderately projective under all four syntactic environments (ranges 39.5–53.5% and 50.5–66.0% for *again* and CSV, respectively).

RoBERTa and DeBERTa judge manner adverbs and comparatives as non-projective out of the negative sentence and counterfactual conditional antecedent (e.g., *The professor did not set the dish on the table quietly.* and *The teacher did not read the letter better than the director.*, respectively). The projectivity of these conditions is below the 33.3% chance level (ranges 8.5–18% and 5.5–23.5% for RoBERTa and DeBERTa, respectively). These results contrast with those of humans as they judge these conditions as moderately projective (in the range 58.3–70.7%).

Finally, RoBERTa and DeBERTa judge triggers as highly projective in the question condition similar to active sentence condition with exceptions

such as manner adverbs (e.g., *Did the boy set the dish on the table calmly?*) and factive predicates (e.g., *Did the assistant remember putting the book on the shelf?*), which indicates that they process these two environments similarly. Again, this result contrasts with human results. Figure 4 shows that human results of the question conditions are different from those of the active sentence conditions. As shown by the difference in the mean projectivity (94.2% and 76.7% for the active sentence and the polar question, respectively), humans assign the question condition relatively low projectivity compared to the active sentence condition, indicating the variability of projection out of the polar question. The high projectivity of the model in the polar question condition can be because of annotation artifacts such as the combination of the lexical overlap bias and the negation bias (Gururangan et al., 2018). Models might label the positive and negative hypothesis as entailment and contradiction, respectively.

To summarize, although RoBERTa and DeBERTa judge most conditions as projective similar to humans, they take some as less projective compared to humans, which indicates that the models' performance on projectivity of presuppositions does not mirror human performance.

6 Discussion

This study investigates whether projectivity of presupposition depends on the combination of various triggers and environments under which they are embedded, and whether the models' performance reflects any variable projection behaviors shown by humans.

By analyzing our intuitive judgment data, we find that humans take all presupposition patterns as projective with some variability. For instance, presuppositions of manner adverbs are less likely to project out of the negative sentence (e.g., *The kid did not bet $100 on the race quickly.*) (58.3%), polar question (e.g., *Did the stranger burst into the room angrily?*) (66.6%), and counterfactual conditional (e.g., *If the woman had slit the envelope easily, the director would have upset the boat.*) (62.0%) than the active sentence (e.g., *The boy read the letter quietly.*) (98.2%). The variable projection patterns shown by human results suggest that it is necessary to look at each trigger and syntactic environment in detail to investigate variable projectivity of presupposition, as many previous studies (Tonhauser et al., 2018; Degen and Tonhauser, 2021b,a) do in the domain of clause-embedding predicates.

Unlike humans, models take some presupposition patterns as non-projective. For instance, RoBERTa and DeBERTa judge manner adverbs as non-projective out of the negative sentence (8.5% and 14.5% for RoBERTa and DeBERTa, respectively) and counterfactual conditional antecedent (17.0% and 5.5%). This result suggests that models cannot make human-like pragmatic inferences and that it is necessary for researchers working on investigation of the language inference ability of models to use a wide range of trigger and environment types.

However, as seen from the results of our experiments, transformer-based models still achieve human-like performance in most cases, which BOW and InferSent models do not, highlighting their sophisticated linguistic performance. It cannot be overstated that the unprecedented advancement of models allows us to look in depth at their ability to process language, as Pavlick (2022) notes.

One limitation of this study is that it does not take into account factors such as context, a lexical item, prior beliefs, and the social identify of the speaker. As shown by previous studies (Karttunen, 1971; Simons, 2001; Tonhauser et al., 2018; Degen and Tonhauser, 2021b), these factors can affect the projectivity of presuppositions. The future work should address the question whether models are sensitive to these factors in the same manner as humans.

7 Conclusion

This paper investigates whether there is variability in the projectivity of presuppositions, depending on triggers and environments, and whether linguistic models mirror humans in processing any variable presupposition behaviors. We create the template-based dataset consisting of 9,800 sentences crossed with six presupposition triggers and four syntactic environments. Using this dataset, we find that presuppositions always project with some variability and models take some presupposition patterns as non-projective unlike humans. Our result suggests that it cannot be simply assumed that presuppositions are always projective and researchers working on model evaluation and dataset creation need to use various types of triggers and environments so that they can investigate models' ability to process presuppositions in detail.

References

David I. Beaver. 1997. Presupposition. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of logic and language*, pages 939–1008. MIT Press and North-Holland, Cambridge, MA and Amsterdam.

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, Lisbon, Portugal. Association for Computational Linguistics.

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, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. In *Sinn und Bedeutung 23*, volume 2, pages 107–124.

Judith Degen and Judith Tonhauser. 2021a. Are there factive predicates? an empirical investigation. *LingBuzz*.

Judith Degen and Judith Tonhauser. 2021b. Prior beliefs modulate projection. *Open Mind*, 5:59–70.

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.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Irene Heim. 1983. On the conversational basis of some presuppositions. In *Proceedings of the 2nd West Coast Conference on Formal Linguistics*, pages 114–125, Stanford, CA. Stanford Linguistics Association.

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*, pages 8690–8705, Online. Association for Computational Linguistics.

Lauri Karttunen. 1971. Some observations on factivity. *Papers in Linguistics*, 4:55–69.

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

Steven C. Levinson. 1983. *Pragmatics*. Cambridge University Press, Cambridge, UK.

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*.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A corpus of naturally-occurring presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366, Online. Association for Computational Linguistics.

Ellie Pavlick. 2022. Semantic structure in deep learning. *Annual Review of Linguistics*, 8.

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.

Christopher Potts. 2015. Presupposition and implicature. In Shalom Lappin and Chris Fox, editors, *The handbook of contemporary semantic theory*, volume 2, pages 168–202. Wiley-Blackwell, Oxford, UK.

Alexis Ross and Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240, Hong Kong, China. Association for Computational Linguistics.

Mandy Simons. 2001. On the conversational basis of some presuppositions. In *Proceedings of Semantics and Linguistics Theory XI*, pages 431–448, Ithaca, NY. CLC Publications.

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

Judith Tonhauser, David I. Beaver, and Judith Degen. 2018. How projective is projective content? gradience in projectivity and at-issueness. *Journal of Semantics*, 35(3):495–542.

Aaron S 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*, pages 221–234.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hitomi Yanaka and Koji Mineshima. 2021. Assessing the generalization capacity of pre-trained language models through Japanese adversarial natural language inference. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 337–349, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Templates

Tables 5–8 contain templates of premises and hypotheses with examples for each trigger crossed with each entailment-canceling environment.

B Crowdsourcing Experiment

Before the experiment, each participant is asked to read a written instruction about the NLI task carefully. During the experiment, the following instruction is presented on a screen: 'Select the response based on how likely you think the second statement is to be true, using the information in the first statement and your background knowledge about how the world works. If you think that the second statement is true, click Entailment. If you think that it is false, select Contradiction. If you are not sure, select Neutral.' This instruction is adopted from Parrish et al. (2021) and modified according to our experiment.

We eliminate participants whose mean accuracy on the control conditions is less than or equal to 90%. The control results include results of active sentence, negative sentence, and counterfactual conditional conditions. The polar question control condition is not included in the mean calculation, because its mean accuracy is around chance (36.0% over the chance level 33.3%).

Type	Templates	Examples
Again	P: The N VP again. $\rightarrow H_1$: The N had VP before. $\not\rightarrow H_2$: The N had not VP before.	The doctor cut the tree again. $\rightarrow H_1$: The doctor had cut the tree before. $\not\rightarrow H_2$: The doctor had not cut the tree before.
Manner adverbs	P: The N VP MADV. $\rightarrow H_1$: The N VP. $\not\rightarrow H_2$: The N did not VP.	The doctor cut the tree slowly. $\rightarrow H_1$: The doctor cut the tree. $\not\rightarrow H_2$: The doctor did not cut the tree.
Comparatives	P: The N_1 VP ADVer than N2. $\rightarrow H_1$: The N_1 VP. $\not\rightarrow H_2$: The N_1 did not VP.	The doctor cut the tree better than the singer. $\rightarrow H_1$: The doctor cut the tree. $\not\rightarrow H_2$: The doctor did not cut the tree.
Temporal adverbs	P: The N VP_1 TADV VP_2ing. $\rightarrow H_1$: The N VP_2. $\not\rightarrow H_2$: The N did not VP_2.	The doctor cut the tree before spreading the rumor. $\rightarrow H_1$: The doctor spread the rumor. $\not\rightarrow H_2$: The doctor did not spread the rumor.
Change-of-state verbs	P: The N CSV VPing. $\rightarrow H_1$: The N had been VPing. $\not\rightarrow H_2$: The N had not been VPing.	The doctor stopped cutting the tree. $\rightarrow H_1$: The doctor had been cutting the tree. $\not\rightarrow H_2$: The doctor had not been cutting the tree.
Factive verbs	P: The N Factive VPing. $\rightarrow H_1$: The N VP. $\not\rightarrow H_2$: The N did not VP.	The doctor regretted cutting the tree. $\rightarrow H_1$: The doctor cut the tree. $\not\rightarrow H_2$: The doctor cut the tree.

Table 5: Templates for presupposition triggers in active sentences.

Type	Templates	Examples
Again	P: The N did not VP again. $\rightarrow H_1$: The N had VP before. $\not\rightarrow H_2$: The N had not VP before.	The doctor did not cut the tree again. $\rightarrow H_1$: The doctor had cut the tree before. $\not\rightarrow H_2$: The doctor had not cut the tree before.
Manner adverbs	P: The N did not VP MADV. $\rightarrow H_1$: The N VP. $\not\rightarrow H_2$: The N did not VP.	The doctor did not cut the tree slowly. $\rightarrow H_1$: The doctor cut the tree. $\not\rightarrow H_2$: The doctor did not cut the tree.
Comparatives	P: The N_1 did not VP ADVer than N2. $\rightarrow H_1$: The N_1 VP. $\not\rightarrow H_2$: The N_1 did not VP.	The doctor did not cut the tree better than the singer. $\rightarrow H_1$: The doctor cut the tree. $\not\rightarrow H_2$: The doctor did not cut the tree.
Temporal adverbs	P: The N did not VP_1 TADV VP_2ing. $\rightarrow H_1$: The N VP_2. $\not\rightarrow H_2$: The N did not VP_2.	The doctor did not cut the tree before spreading the rumor. $\rightarrow H_1$: The doctor spread the rumor. $\not\rightarrow H_2$: The doctor did not spread the rumor.
Change-of-state verbs	P: The N did not CSV VPing. $\rightarrow H_1$: The N had been VPing. $\not\rightarrow H_2$: The N had not been VPing.	The doctor did not stop cutting the tree. $\rightarrow H_1$: The doctor had been cutting the tree. $\not\rightarrow H_2$: The doctor had not been cutting the tree.
Factive verbs	P: The N did not Factive VPing. $\rightarrow H_1$: The N VP. $\not\rightarrow H_2$: The N did not VP.	The doctor did not regret cutting the tree. $\rightarrow H_1$: The doctor cut the tree. $\not\rightarrow H_2$: The doctor did not cut the tree.

Table 6: Templates for presupposition triggers in a negative sentence.

Type	Templates	Examples
Again	P: Did the N VP again? $\rightarrow H_1$: The N had VP before. $\not\rightarrow H_2$: The N had not VP before.	Did the doctor cut the tree again? $\rightarrow H_1$: The doctor had cut the tree before. $\not\rightarrow H_2$: The doctor had not cut the tree before.
Manner adverbs	P: Did the N VP MADV? $\rightarrow H_1$: The N VP. $\not\rightarrow H_2$: The N did not VP.	Did the doctor cut the tree slowly? $\rightarrow H_1$: The doctor cut the tree. $\not\rightarrow H_2$: The doctor did not cut the tree.
Comparatives	P: Did the N_1 VP ADVer than N2? $\rightarrow H_1$: The N_1 VP. $\not\rightarrow H_2$: The N_1 did not VP.	Did the doctor cut the tree better than the singer? $\rightarrow H_1$: The doctor cut the tree. $\not\rightarrow H_2$: The doctor did not cut the tree.
Temporal adverbs	P: Did the N VP_1 TADV VP_2ing? $\rightarrow H_1$: The N VP_2. $\not\rightarrow H_2$: The N did not VP_2.	Did the doctor cut the tree before spreading the rumor? $\rightarrow H_1$: The doctor spread the rumor. $\not\rightarrow H_2$: The doctor did not spread the rumor.
Change-of-state verbs	P: Did the N CSV VPing? $\rightarrow H_1$: The N had been VPing. $\not\rightarrow H_2$: The N had not been VPing.	Did the doctor stop cutting the tree? $\rightarrow H_1$: The doctor had been cutting the tree. $\not\rightarrow H_2$: The doctor had not been cutting the tree.
Factive verbs	P: Did the N Factive VPing? $\rightarrow H_1$: The N VP. $\not\rightarrow H_2$: The N did not VP.	Did the doctor stop cutting the tree? $\rightarrow H_1$: The doctor cut the tree. $\not\rightarrow H_2$: The doctor did not cut the tree.

Table 7: Templates for presupposition triggers in a yes-no question.

Type	Templates	Examples
Again	P: If the N_1 had VP again, the N2 would have VP_2. $\rightarrow H_1$: The N_1 had VP_1 before. $\not\rightarrow H_2$: The N_1 had not VP_1 before.	If the doctor had cut the tree again, the singer could have spread the news. $\rightarrow H_1$: The doctor had cut the tree before. $\not\rightarrow H_2$: The doctor had not cut the tree before.
Manner adverbs	P: If the N_1 VP_1 MADV, the N2 would have VP_2. $\rightarrow H_1$: The N_1 VP_1. $\not\rightarrow H_2$: The N_1 did not VP_1.	If the doctor cut the tree slowly, the singer could have spread the news. $\rightarrow H_1$: The doctor cut the tree. $\not\rightarrow H_2$: The doctor did not cut the tree.
Comparatives	P: If the N_1 had VP_1 ADVer than N3, the N2 would have VP_2. $\rightarrow H_1$: The N_1 VP_1. $\not\rightarrow H_2$: The N_1 did not VP_1.	If the doctor had cut the tree better than the singer, the artist could have spread the news. $\rightarrow H_1$: The doctor cut the tree. $\not\rightarrow H_2$: The doctor did not cut the tree.
Temporal adverbs	P: If the N had VP_1 TADV VP_2ing, the N2 would have VP3. $\rightarrow H_1$: The N_1 VP_2. $\not\rightarrow H_2$: The N_1 did not VP_2.	If the doctor had cut the tree before spreading the news, the singer could have burst into the room. $\rightarrow H_1$: The doctor spread the rumor. $\not\rightarrow H_2$: The doctor did not spread the rumor.
Change-of-state verbs	P: If the N_1 CSV VP_1ing, the N2 would have VP_2. $\rightarrow H_1$: The N_1 had been VP_1ing. $\not\rightarrow H_2$: The N_1 had not been VP_1ing.	If the doctor had stopped cutting the tree, the singer could have spread the rumor. $\rightarrow H_1$: The doctor had been cutting the tree. $\not\rightarrow H_2$: The doctor had not been cutting the tree.
Factive verbs	P: If the N_1 Factive VP_1ing, the N2 would have VP_2. $\rightarrow H_1$: The N_1 VP_1. $\not\rightarrow H_2$: The N_1 did not VP_1.	If the doctor had stopped cutting the tree, the singer could have spread the rumor. $\rightarrow H_1$: The doctor cut the tree. $\not\rightarrow H_2$: The doctor did not cut the tree.

Table 8: Templates for presupposition triggers in an antecedent of a counterfactual conditional.