CoNLL 2023

*daiasami@udel.edu

PROPRES: Investigating the projectivity of presupposition with various triggers and environments

*Daiki Asami (University of Delaware) Saku Sugawara (National Institute of Informatics)

Abstract

- Our human evaluation shows that projectivity of presupposition can vary depending on triggers and environments.
- The best performed model, DeBERTa, does not capture the variable projectivity observed in IMPPRES (Jeretic+ 2021)
- We introduce a new dataset PROPRES ahd shows that DeBERTa does not capture the variable projectivity observed in it.
- Studies on pragmatic inferences should take extra care of the human judgment variability and combination of linguistic items.

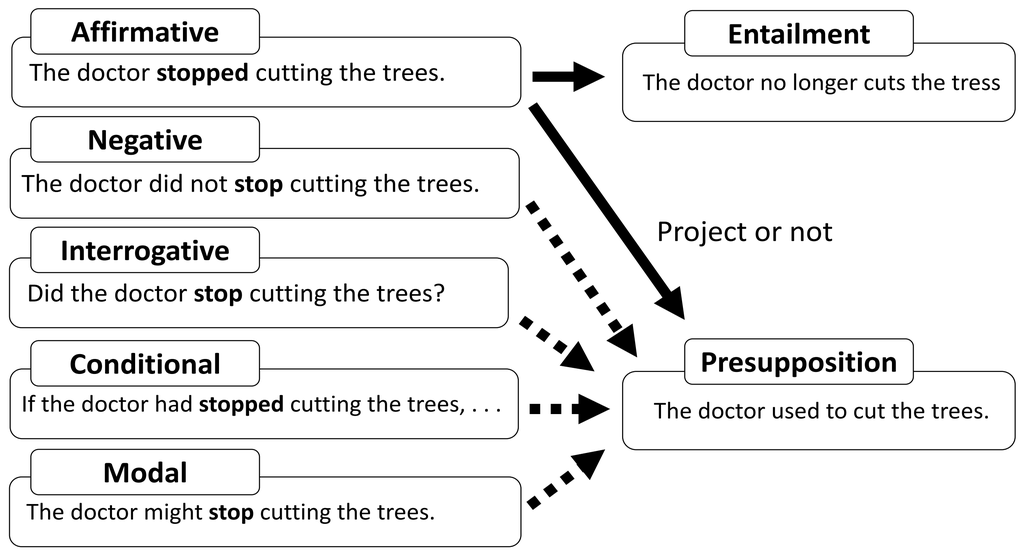
Background and Motivation

Projectivity of Presupposition in Linguistics

- Presupposition is introduced by a *presupposition trigger* (e.g., 'again').
- Presupposition *projects* out of entailment canceling environments (e.g., negation).
- Q. Does the projectivity depends on the combination of triggers and environments?
- To tackle this question, we collect human evaluation data on a previous dataset and our dataset.

Presupposition in NLI studies

- Jeretic+ (2020) probe language models' performance on projectivity with IMPPRES but conduct no human evaluation. Q. Can models capture variable projectivity exhibited by humans?
- Parrish+ (2021) collect human evaluation data but use only negation as an entailment-canceling environment. Q. Can models capture projectivity out of other environments?
- We resolve these issues by both conducting human evaluation data and using various entailment-canceling environments.



Experiment 1: Reevaluation of IMPPRES (**Imp**recature and **pres**upposition: Jeretic+ (2021))

IMPRESS

9 presupposition triggers* 5 environments

Trigger	Example	Presupposition
All N	All four waiters that bothered Paul telephoned.	Exactly four waiters telephoned.
Both	Both people that hoped to move have married.	Exactly two people have married.
Change of state verb	Marie was leaving.	Marie was here.
Cleft existence	It is Margaret that forgot Dan.	Someone forgot Dan.
Cleft uniqueness	It is Donna who studied.	Exactly one person studied.
Only	The pasta only annoys Roger.	The pasta annoys Roger.
Possessive definites	The boy's rugs did look like these prints.	The boy has rugs.
Possessive uniqueness	Maria's apple that ripened annoys the boy.	Maria has exactly one apple that ripened.
Question	Bob learns how Rachel approaches Melanie.	Rachel approaches Melanie.

Environment	Example
Affirmative sentence	All four waiters that bothered Paul telephoned.
Negation	All four waiters that bothered Paul did not telephone.
Interrogative	Did all four waiters that bothered Paul telephone?
Conditional	If all four waiters that bothered Paul telephoned, it's okay.
Modal	All four waiters that bothered Paul might telephone.

Human Evaluation & Models

Human evaluation: 9.4 labels for each item on average on AMT.
Models: RoBERTa-base & DeBERTa-large finetuned on MNLI.

Reusults

Affirmative sentences:

- In an affirmative sentence, presupposition equals entailment. Our evaluation metric is accuracy.

Humans:

1. CoS (66.3%), 2. Cleft unique. (74.1%), 3. Possessive unique. (71.9%) (other triggers (acc. > 87.3%))
   e.g., CoS : Omar is hiding Ben. → Ben was out in the open.
   Cleft unique.: It is that doctor who left.→ Exactly one person left.
   Possess. unique.: Tom's car that broke bored this committee. → Tom has exactly one car that broke.

- The judgment of these data is not as robust as theoreticians assume.

RoBERTa & DeBERTa:

4. *All* N (71.0% & 89.5%), 5. *Both* (39.0% & 49.0%)
   e.g., *All* N: all four men that departed telephoned. → Exactly four men departed.
   Both: both guys who ran jumped. → Exactly two guys ran.

- We remove these 5 triggers from the following analysis.

Entailment-canceling environments:

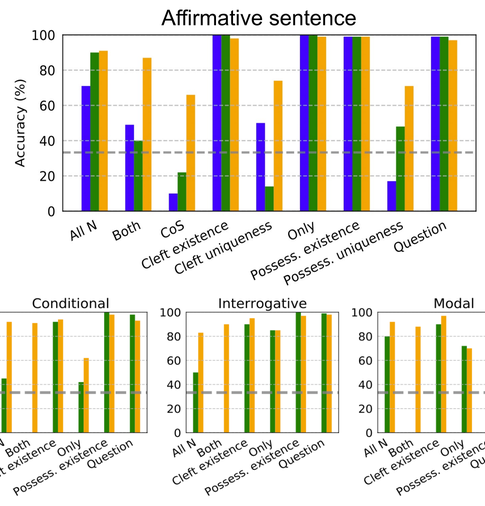
- We use projectivity instead of accuracy.

All N in conditional (91.8% vs. 45.0%)
e.g., If all nine actors that left slept, ... → Exactly nine actors left.

All N in interrogative (82.6% vs. 49.5%)
e.g., Did all nine actors that left sleep? → Exactly nine actors left.

Cleft exist. in conditional (89.7% vs. 65.0%)
e.g., If it is Margaret that forgot Dan, … → Someone forgot Dan.

- Humans and DeBERTa show similarity in other conditions.



Experiment 2: PROPRES (**Pro**jectivity of **pres**upposition)

PROPRES

6 triggers*5 environments

Trigger Type	Example Triggers	Example Premise
Iterative	*again*	The assistant split the log **again**.
Aspectual 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 than*, *earlier than*	The assistant split the log **better than** the girl.
Temporal adverb	*before*, *after*, *while*	The assistant split the log **before** bursting into the room.

Environment	Premise	Hypothesis (target and control)
Unembedded	The doctor shed tears again.	Target: The doctor had (not) shed tears before.
Negation	The doctor did not shed tears again.	
Interrogative	Did the doctor shed tears again?	
Conditional	If the doctor had shed tears again, ...	Control: The doctor (did not) shed tears again.
Modal	The doctor might shed tears again.	

Human Evaluation & Models

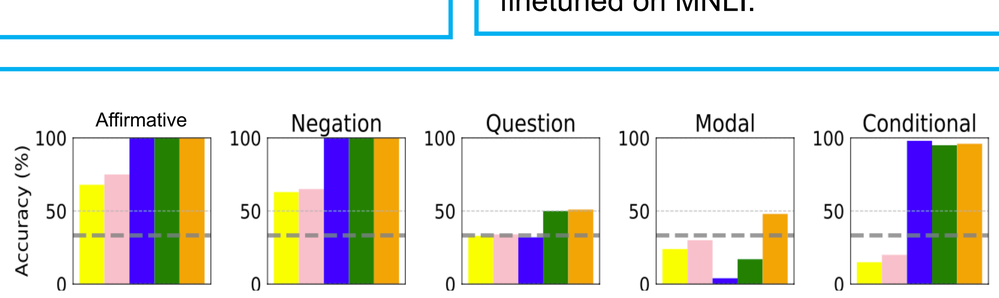
Human evaluation: 56.7 or 9.4 labels on AMT (the difference occcurs because we conduct a human evaluation on a subset of PROPRES in Experiment 1).

Models: BOW & InferSent (baseline)
RoBERTa-base & DeBERTa-large finetuned on MNLI.

Results

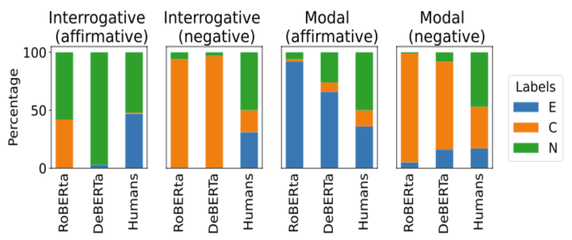
Control conditions:

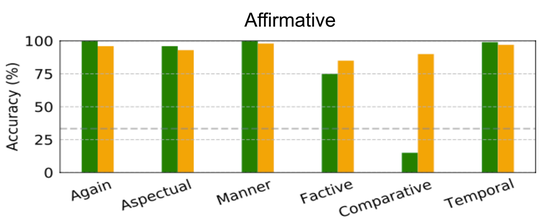
- In control conditions, hypotheses are affirmative or negative versions of their premises.
- They surve as sanity chech in a human evaluation and allow us to check whether models rely on lexical overlap (McCoy+ 2019) or negation heuristics (Gururangan+ 2018).
- Infersent and BOW perform poorly.
- Humans, RoBERTa, and DeBERTa perform well on uneenbedded, negation, and conditional conditions.
- Humans, RoBERTa, and DeBERTa perform poorly on interogative and modal conditions whose correct labels are neutral.



Control interrogative and modal conditions:

- Humans' Interrogative with an affirmative hypothesis: Entailment (46.5%) & Neutral (52.4%)
- They seem ambiguous between yes/no and confirmation questions whose gold label is entailment.
- RoBERTa and DeBERTa do not perform like humans.
- We do not analyze the interrogatives and modals below.





Affirmative sentences:

- Humans and DeBERTa achieve high accuracy.
- DeBERTa performs poorly on the comparaitve (65.0%). e.g., The girl read the letter better than the boy. → The boy read the letter.

Entailment canceling environments.

- Humans show variable projectivity (range 55.1–98.8%)

e.g., Manner adverbs in negation (58.3%) e.g., The man did not hurt others seriously. →The man hurt others seriously.
in interrogative (66.6%) e.g., Did the man hurt others seriously? →The man hurt others seriously.
in conditional (62.0%) e.g., If the man had hurt others seriously, . . . →The man hurt others seriously.
in modal (55.1%) e.g., The man might hurt others seriously. →The man hurt others seriously.
Temporal adverbs in modal (54.7%) e.g., The man might sing after reading. → The man read.

- DeBERTa does not behave like humans in some cases e.g., manner adverbs in negation (8.5%) and conditional (14.0%).