Inference is Everything: Recasting Semantic Resources into a Unified Evaluation Framework

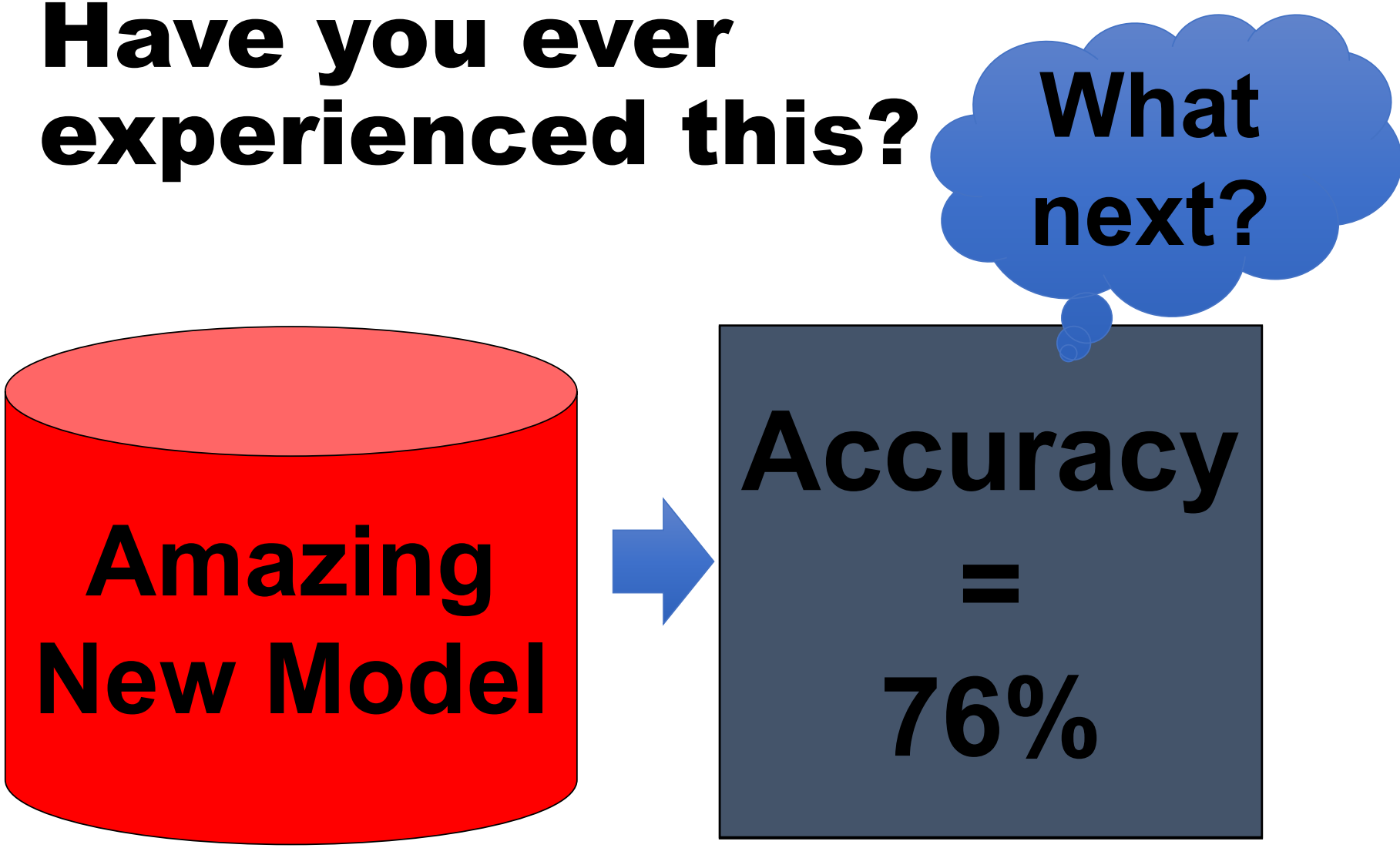
Aaron White (Rochester)

Kevin Duh (JHU)

Pushpendre Rastogi (JHU)

Benjamin Van Durme (JHU)

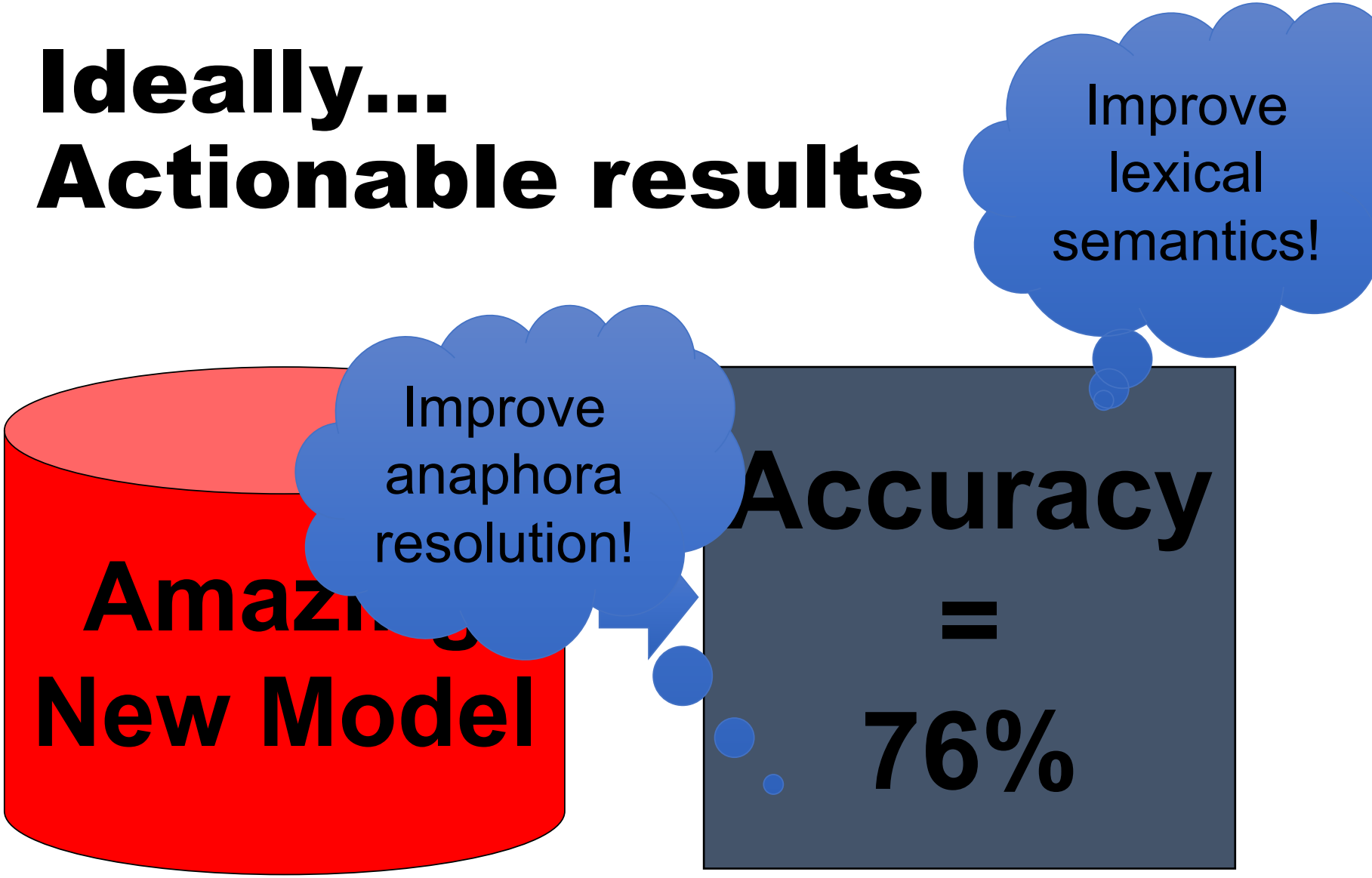
Have you ever experienced this?



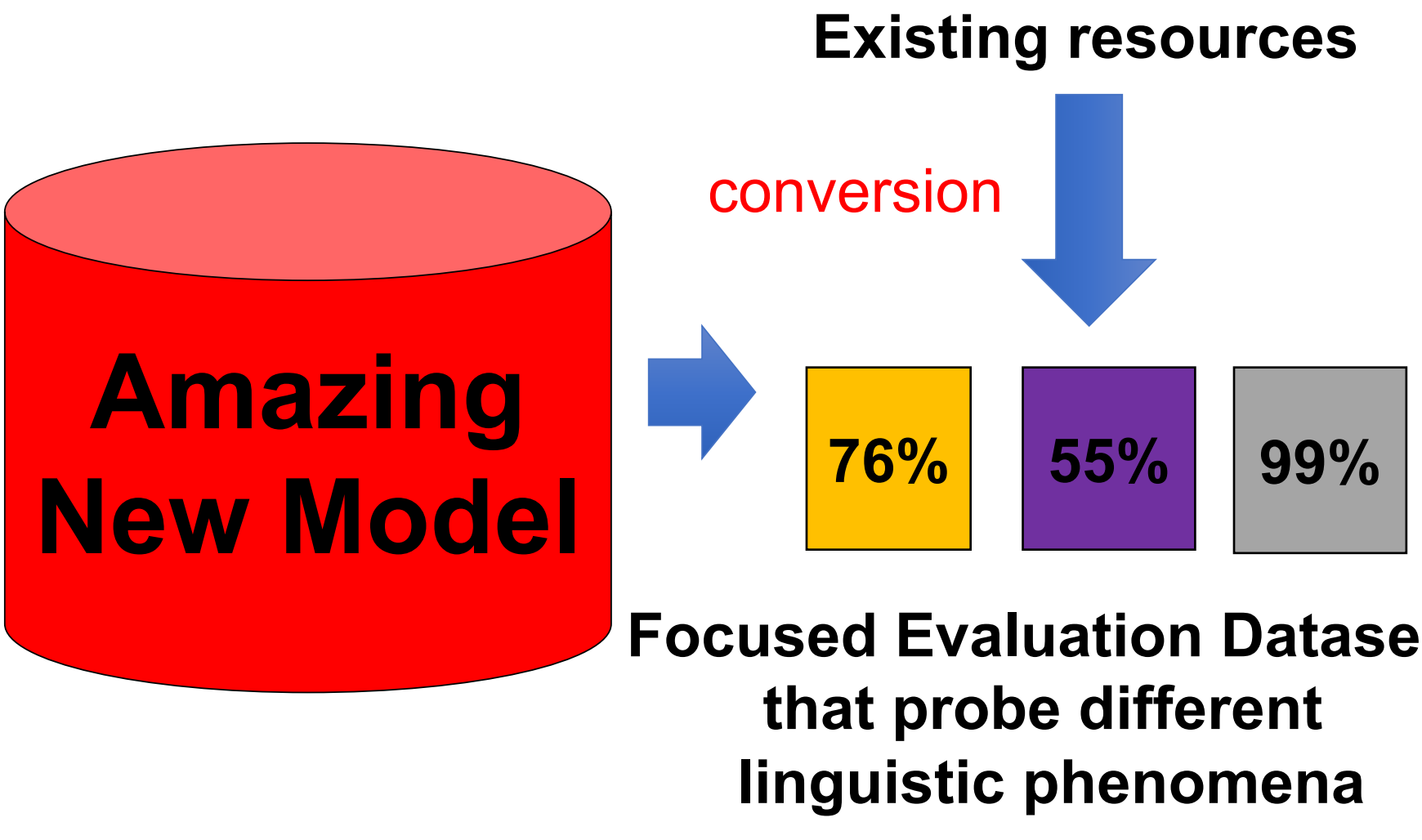
e.g. for Recognizing Textual Entailment (RTE)

e.g. Stanford Natural Language Inference (SNLI) dataset

Ideally…
Actionable results



Idea (for RTE)


Existing resources
conversion
Amazing New Model
76%
55%
99%
Focused Evaluation Datasets that probe different linguistic phenomena

Previous work with similar motivations

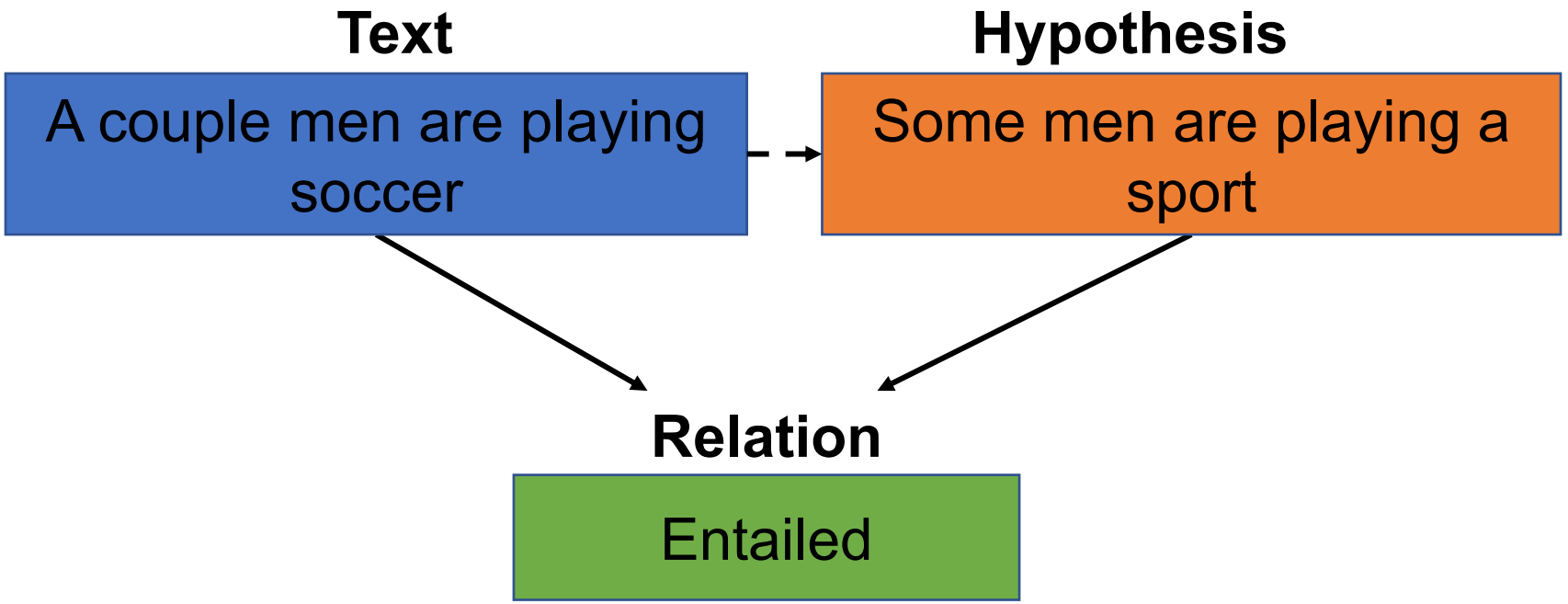
- FraCaS [Cooper et. al. 1996]
  - Manually constructed test suite to probe a range of semantic phenomena
- bAbI [Weston et. al. 2016]
  - Automatically generated test suite to probe different capabilities needed in question answering
- Challenge set for Machine Translation [Isabelle, 2017]
  - Manually constructed reference set to test subject-verb agreement, noun compounds, question syntax, etc.

Outline

1. Motivation
2. Creating focused RTE datasets
3. Case study: debugging neural models

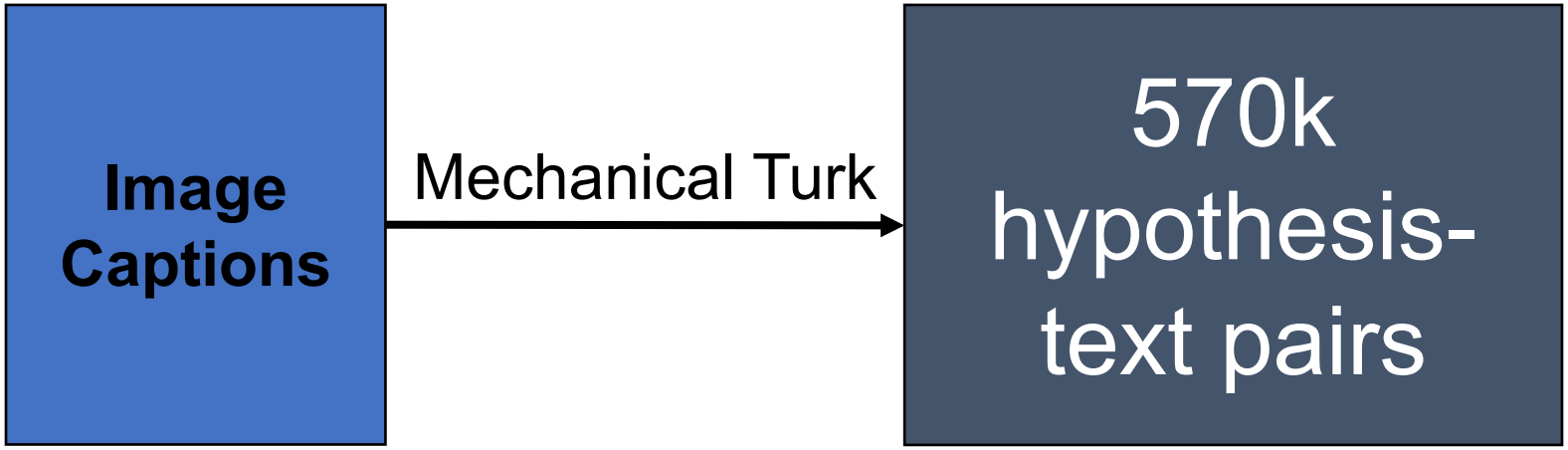
Recognizing Textual Entailment (RTE)

Dagan et al., 2006, 2013; Bar-Haim et al., 2006; Giampiccolo et al., 2007, 2009; Bentivogli et al., 2009, 2010, 2011



Stanford Natural Language Inference data (SNLI)

Bowman et al. 2015



Flickr30k

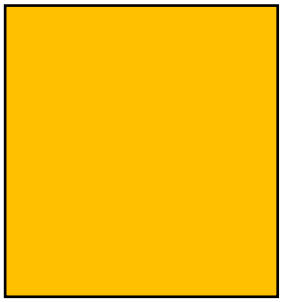
Young et al. 2014

Large-scale data enables training sophisticated models.

But maybe not ideal for evaluation: no fine-grain relations.

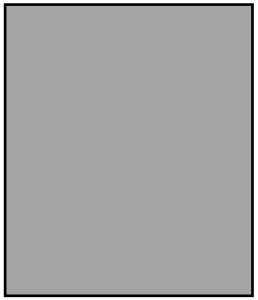
Our contributions

An evaluation framework based on *recasting* existing classification datasets to RTE, e.g.:



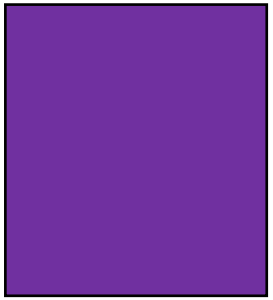
Definite Pronoun Resolution (DPR)

Rahman and Ng 2012



FrameNet Plus (FN+)

Pavlick et al. 2015



Semantic Proto-Roles (SPR)

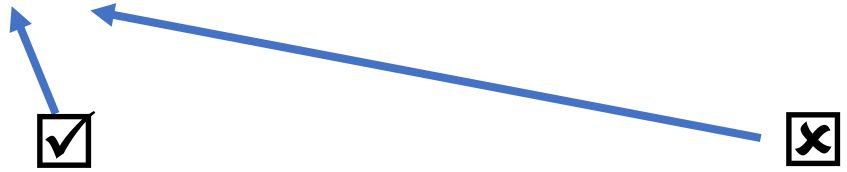
Reisinger et al., 2015

Recasting Definite Pronoun Resolution (DPR) to RTE

Original classification task:

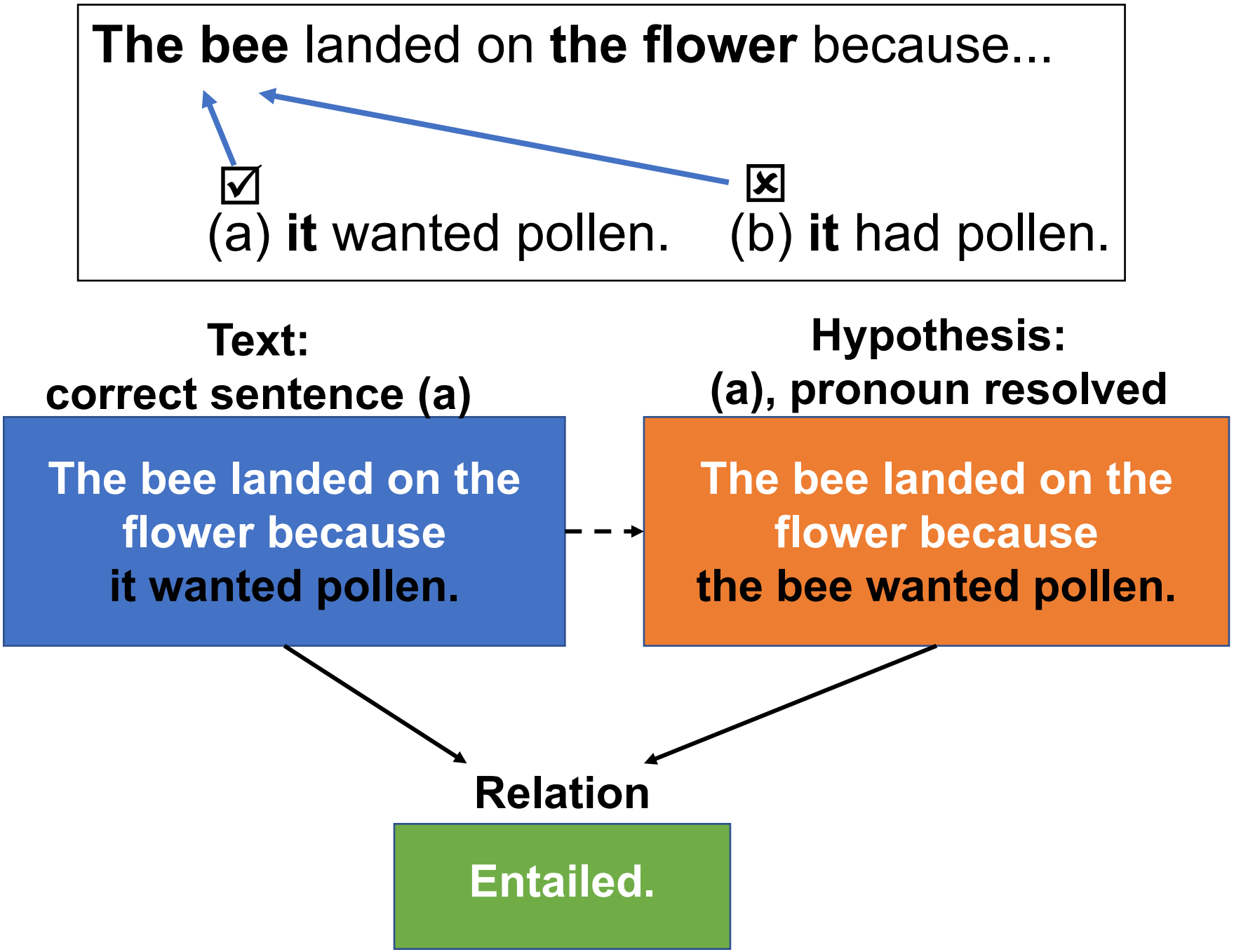
- Map pronoun to coreferential element.
- A step towards the Winograd Challenge

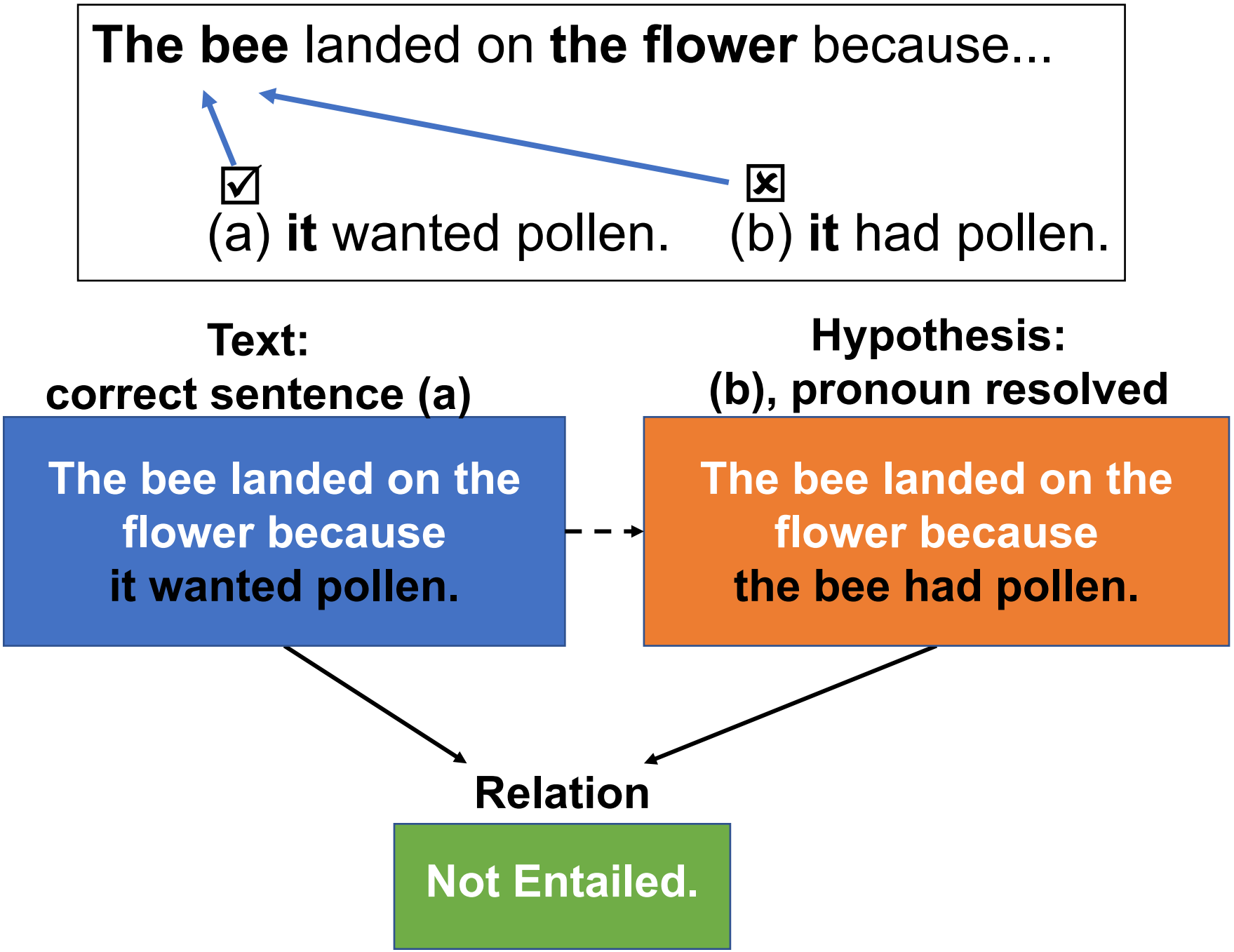
The bee landed on **the flower** because...



☑ (a) **it** wanted pollen.

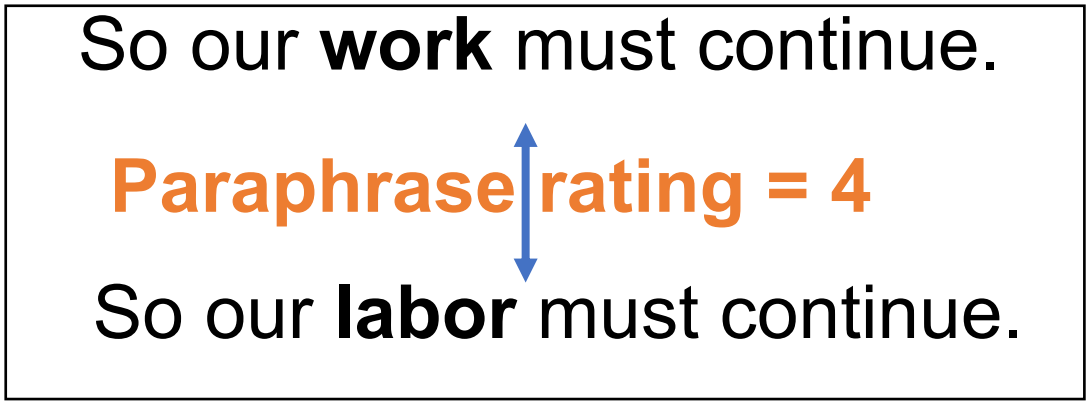
☒ (b) **it** had pollen.


The bee landed on the flower because...
☑
☒
(a) it wanted pollen.
(b) it had pollen.
Text:
correct sentence (a)
Hypothesis:
(a), pronoun resolved
The bee landed on the flower because it wanted pollen.
The bee landed on the flower because the bee wanted pollen.
Relation
Entailed.


The bee landed on the flower because...
☑
☒
(a) it wanted pollen.
(b) it had pollen.
Text:
correct sentence (a)
Hypothesis:
(b), pronoun resolved
The bee landed on the flower because it wanted pollen.
The bee landed on the flower because the bee had pollen.
Relation
Not Entailed.

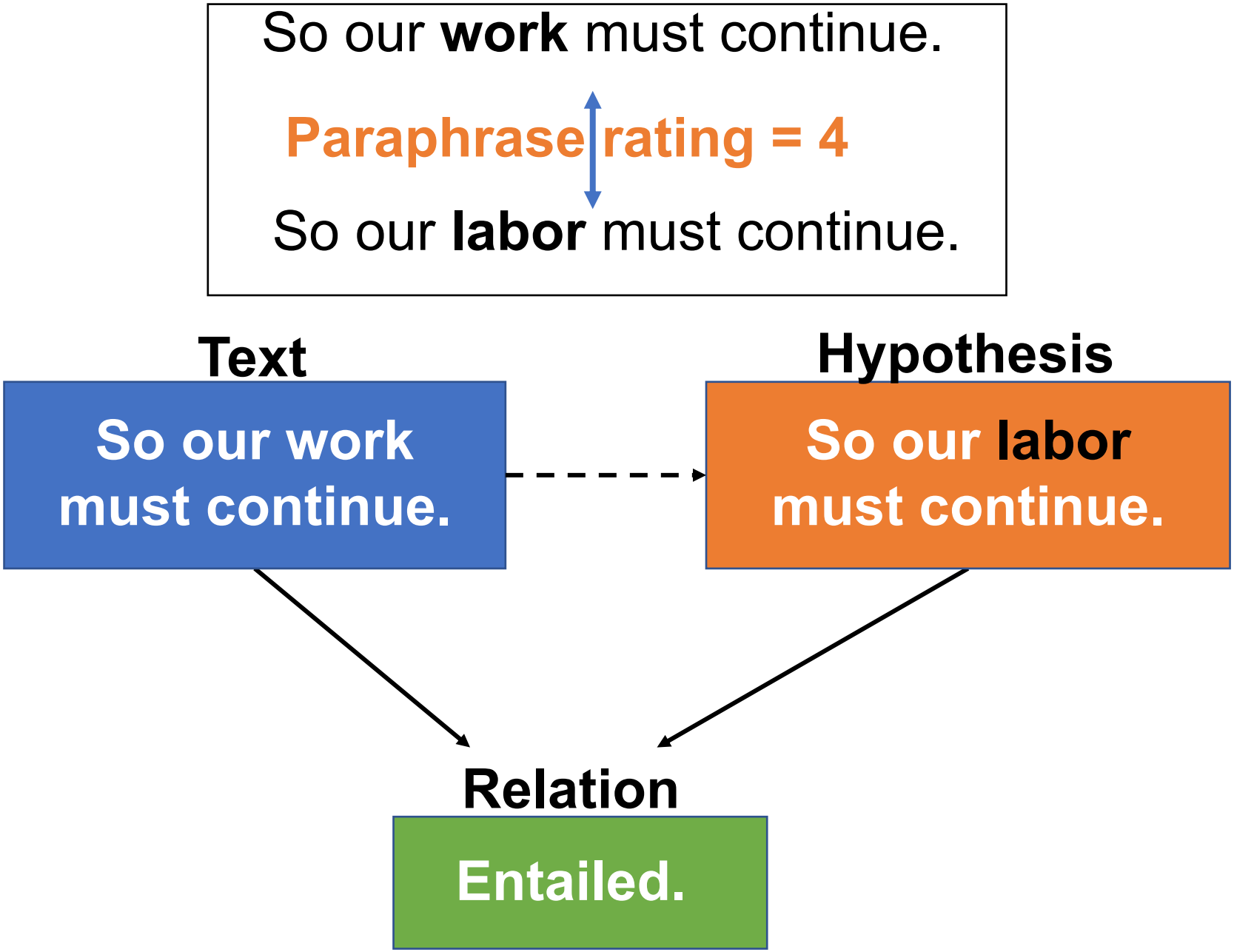
Recasting FrameNet Plus (FN+) to RTE

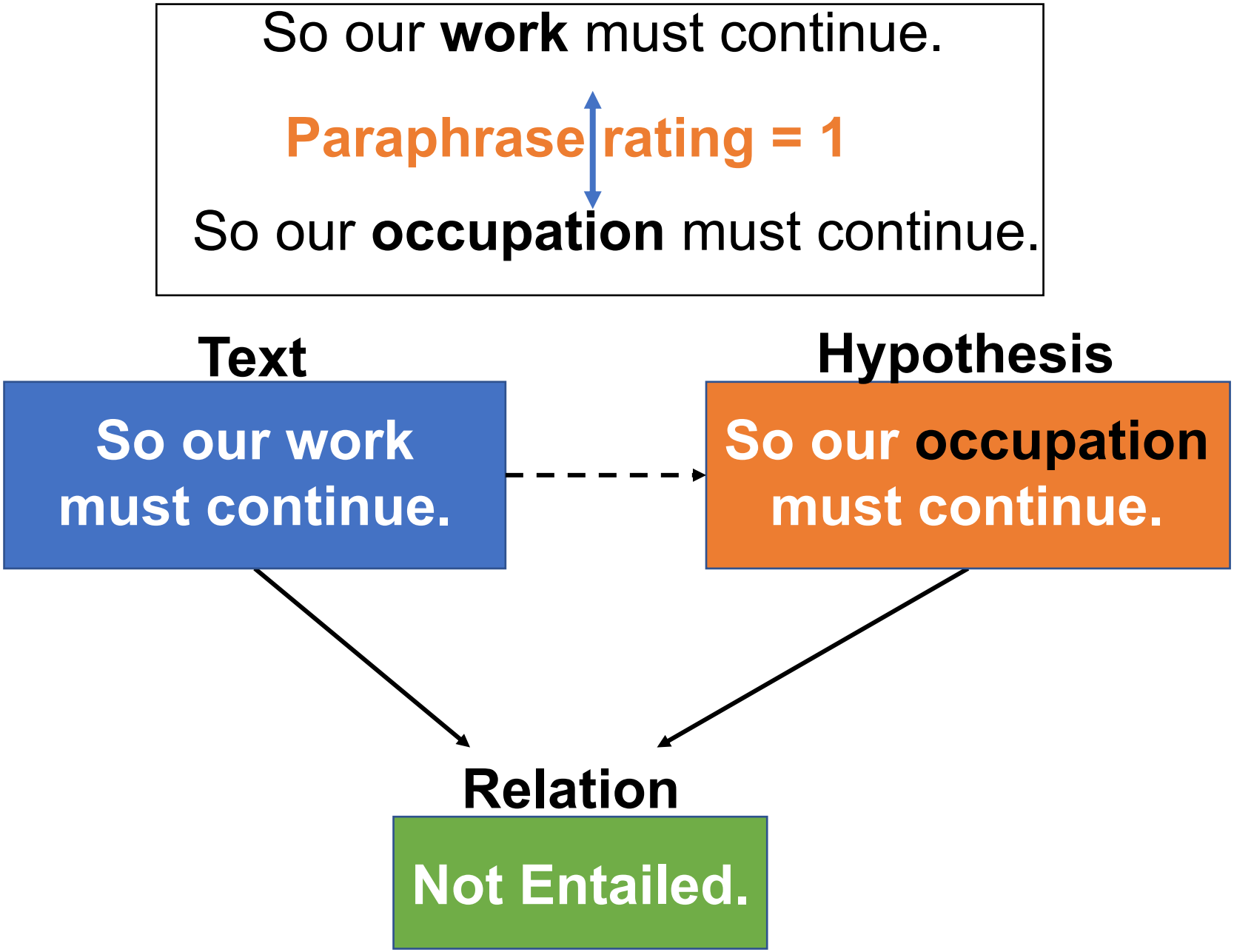
Original data:
- Applied paraphrase to FrameNet triggers
- Turker judged on 5-point scale how much meaning was retained



1-3 rating → Not entailed

4-5 rating → Entailed


So our **work** must continue.
Paraphrase rating = 4
So our **labor** must continue.
Text
So our work must continue.
Hypothesis
So our labor must continue.
Relation
Entailed.


So our **work** must continue.
Paraphrase rating = 1
So our **occupation** must continue.
Text
So our work must continue.
Hypothesis
So our occupation must continue.
Relation
Not Entailed.

Recasting Semantic Proto-Roles (SPR) to RTE

EXAMPLES:

- T: I heard parts of the building above my head cracking
- H: I was aware of being involved in the hearing

- T: UNESCO converted the founding U.N. ideals of individual rights and liberty into peoples' rights
- H: UNESCO existed after the converting stopped

- T: THE IRS delays several deadlines for Hugo's victims
- H: THE IRS caused the delaying to happen.

Semantic Proto-Roles

- What's the number and character of thematic roles in the syntax/semantics interface?
  - AGENT and PATIENT
  - BENEFICIARY? RECIPIENT? Fuzzy boundaries?
- Dowty (1991) introduced Proto-Agent, Proto-Patient fine-grained properties
  - Did the argument change state?
  - Did the argument have volition in the change?

Example Semantic Proto-Role Properties

Role property	How likely or unlikely is it that...
instigation	ARG caused the PRED to happen?
volition	ARG chose to be involved in the PRED?
sentient	ARG was/were sentient?
change of location	ARG changed location during the PRED?
exists as physical	ARG existed as a physical object?
existed before	ARG existed before the PRED began?

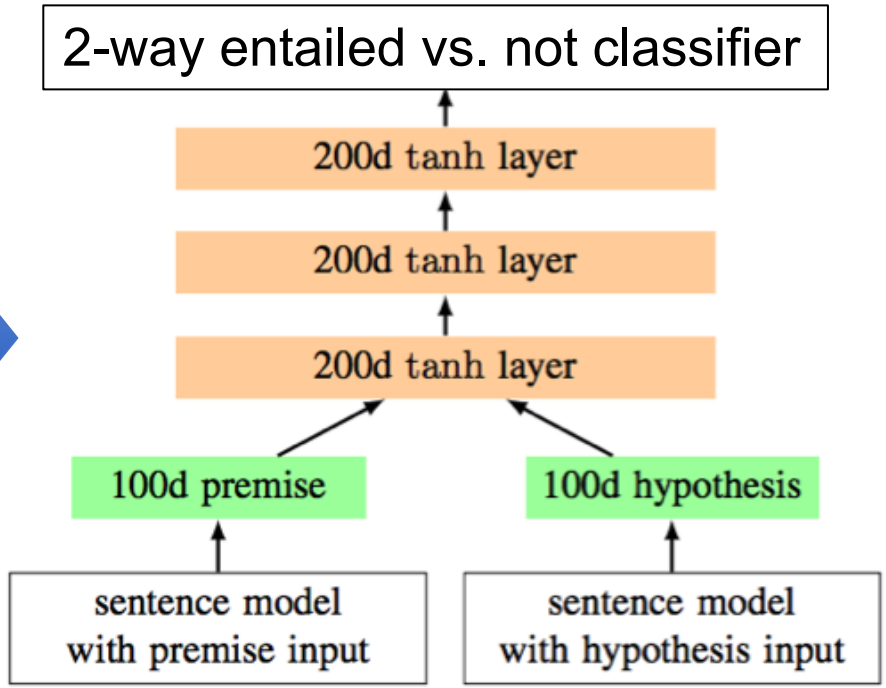
Focused RTE Dataset characteristics

Dataset	Sentences	Label Percentage	
		Entailed	*Not-Entailed*
FN+	154,605	43.45	56.55
SPR	154,607	34.80	65.20
DPR	3,661	49.99	50.01
Total	312,873	39.13	60.87
SNLI†	569,033	33.41	66.59

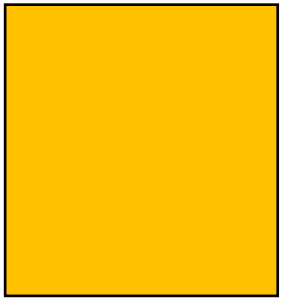
Outline

1. Motivation
2. Creating focused RTE datasets
3. Case study: debugging neural models

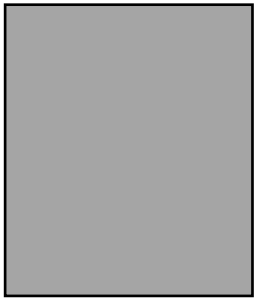
Train on SNLI



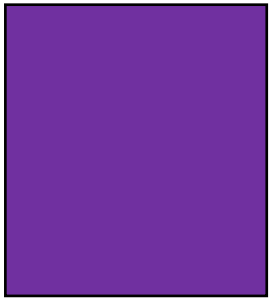
Evaluated on recasted focused RTE datasets:



Definite Pronoun Resolution (DPR)

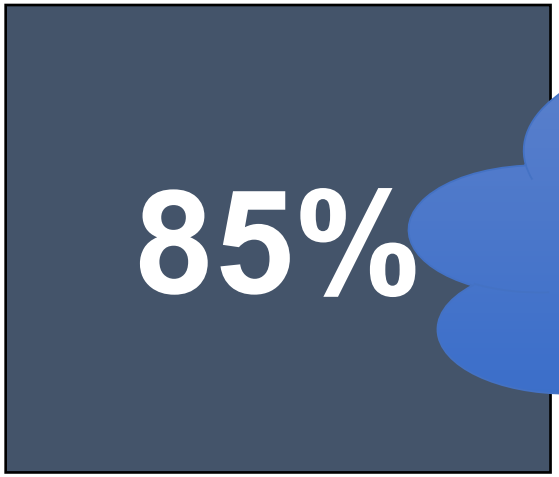


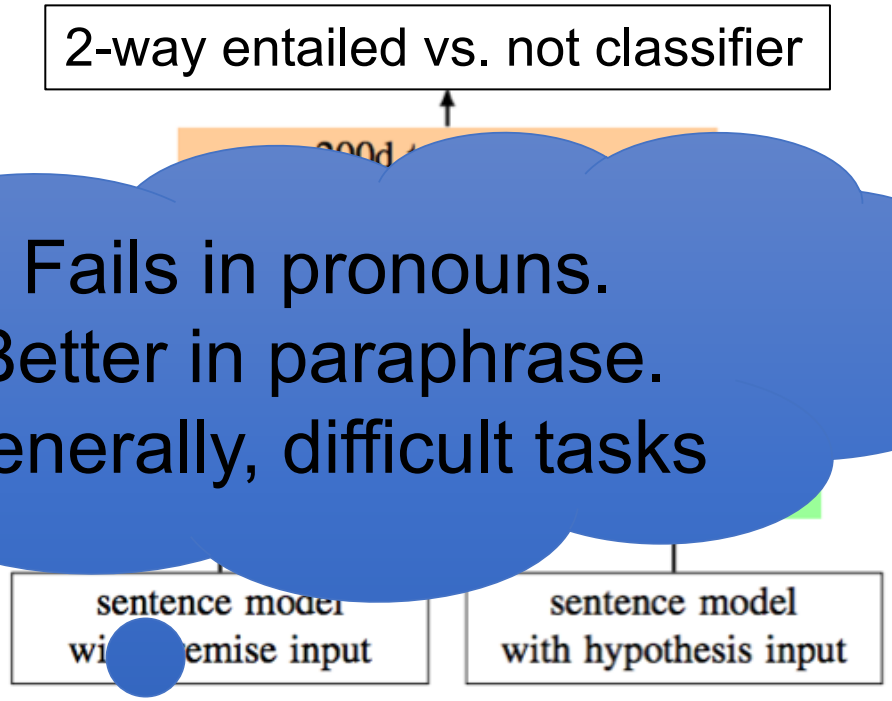
FrameNet Plus (FN+)



Semantic Proto-Roles (SPR)

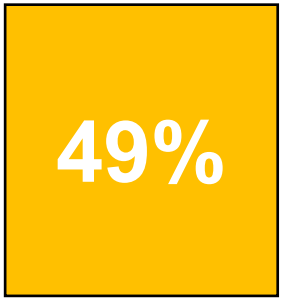
Train on SNLI



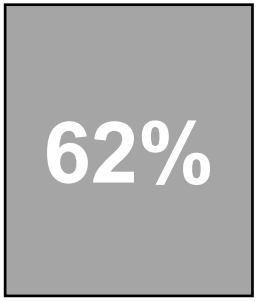


Fails in pronouns.
Better in paraphrase.
Generally, difficult tasks

Evaluated on recasted focused RTE datasets:

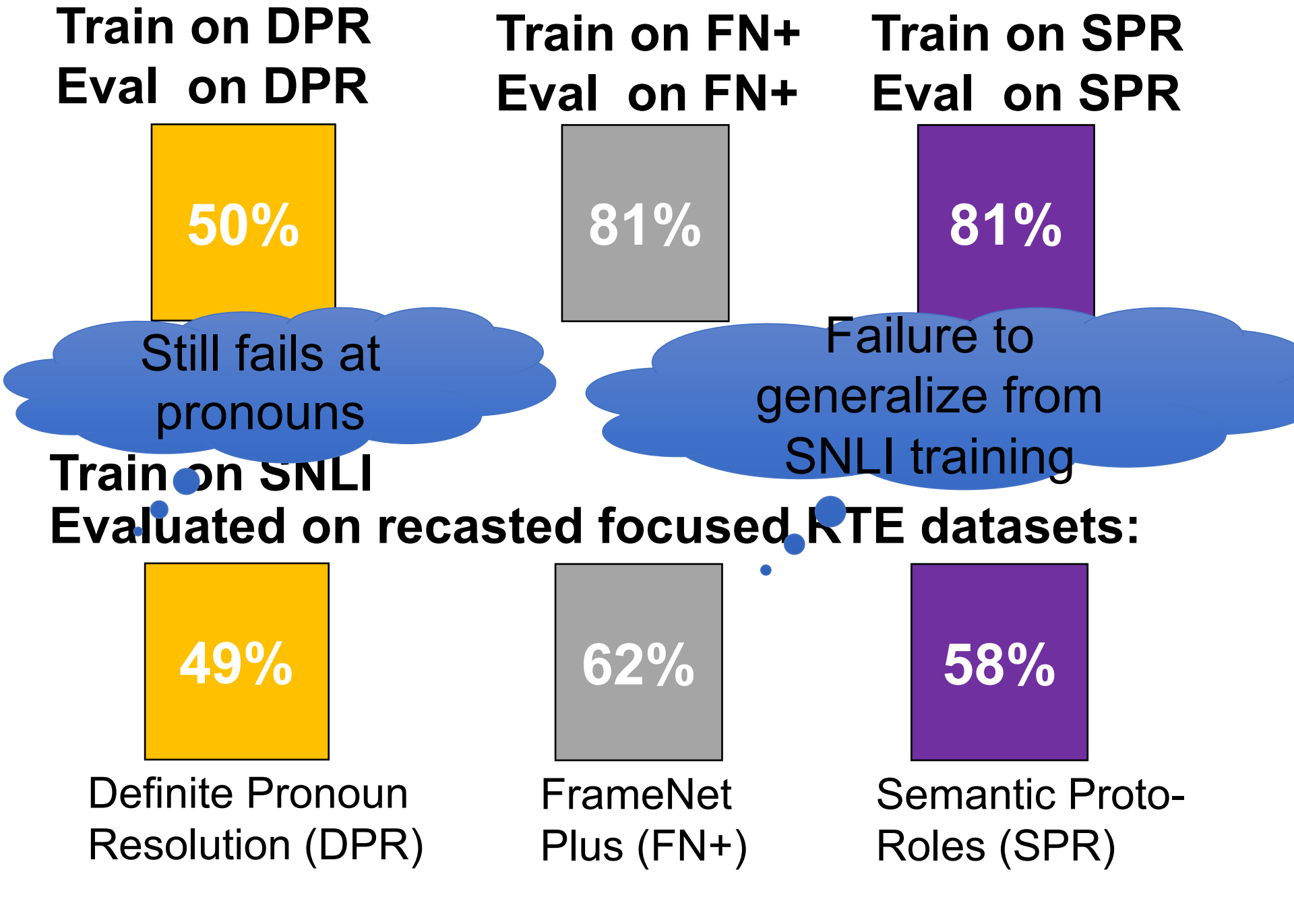


Definite Pronoun Resolution (DPR)

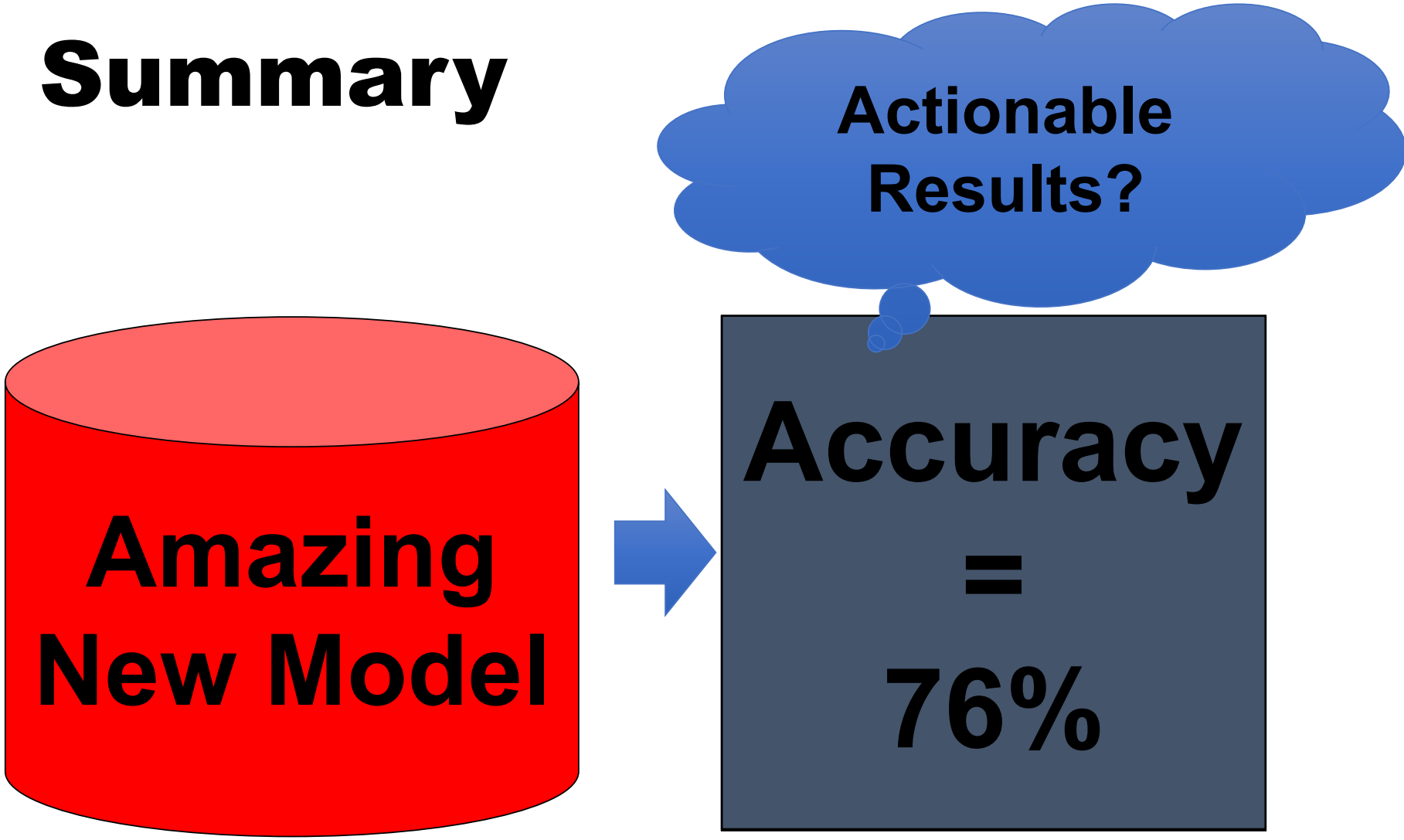


FrameNet Plus (FN+)

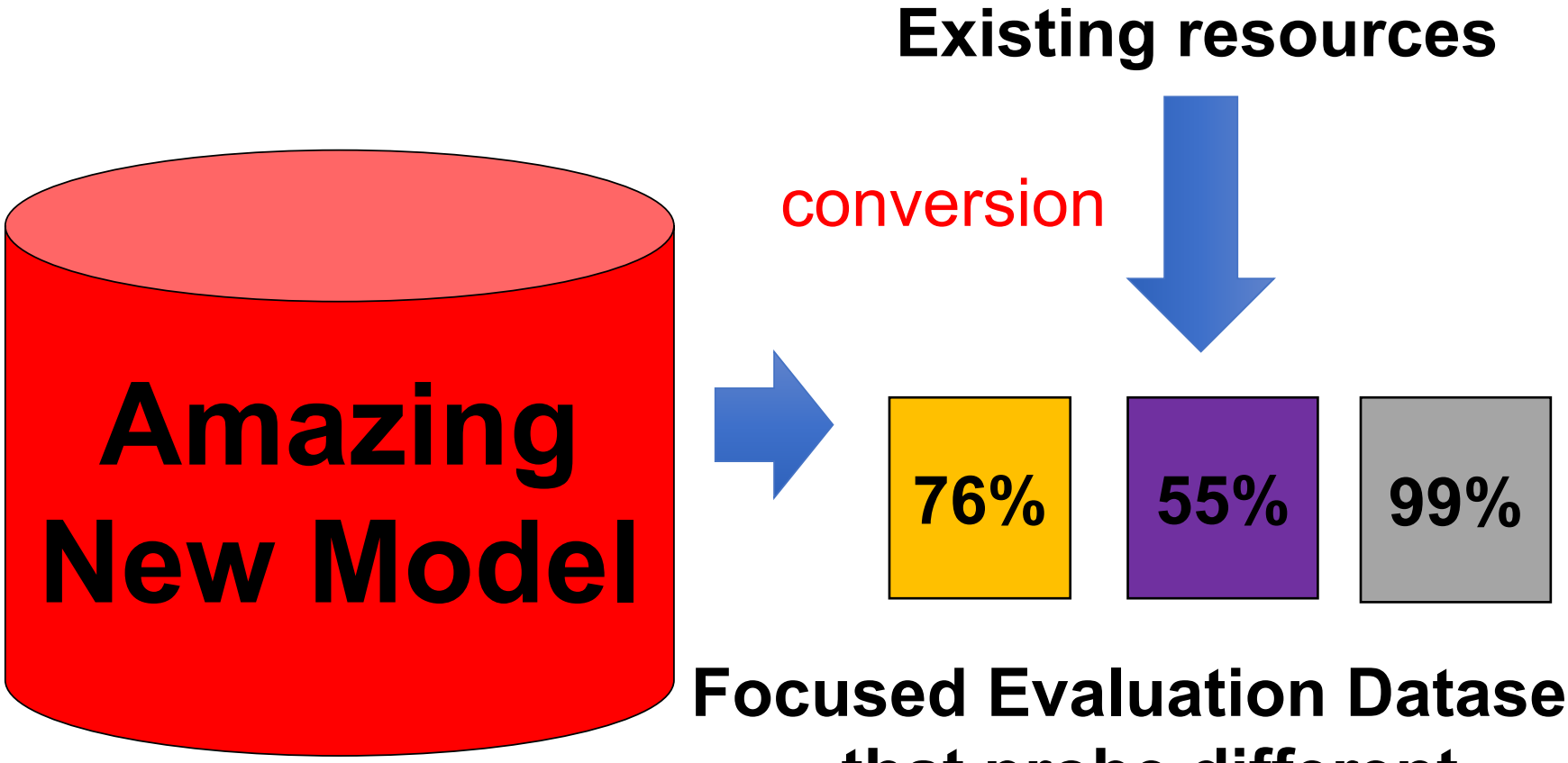
Semantic Proto-Roles (SPR)


Train on DPR
Eval on DPR
50%
Train on FN+
Eval on FN+
81%
Train on SPR
Eval on SPR
81%
Still fails at pronouns
Failure to generalize from SNLI training
Train on SNLI
Evaluated on recasted focused RTE datasets:
49%
62%
58%
Definite Pronoun Resolution (DPR)
FrameNet Plus (FN+)
Semantic Proto-Roles (SPR)

Summary


Actionable Results?
Amazing New Model
Accuracy = 76%
e.g. for Recognizing Textual Entailment (RTE)
e.g. Stanford Natural Language Inference (SNLI) dataset

Summary



Focused Evaluation Datasets that probe different semantic phenomena

(Data available at http://**decomp.net**)

Data Validation

- Manual check of 100 pairs per dataset

Dataset	Accuracy	Grammaticality
FN+	85	77
SPR	94	92
DPR	98	96
SNLI	91	96