Hypothesis Only Models in Natural Language Inference

***SEM 2018**

Adam Poliak, Jason Naradowsky, Aparajita Haldar,
Rachel Rudinger, Benjamin Van Durme

JOHNS HOPKINS UNIVERSITY

Co-Authors

Jason Naradowsky

Aparajita Haldar

Rachel Rudinger

Benjamin Van Durme

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

entailment neutral contradiction

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

entailment neutral contradiction

Johns Hopkins University

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

entailment neutral contradiction

Natural Language Inference

Premise: ***The brown cat ran***

Hypothesis: ***The animal moved***

entailment neutral contradiction

Hypothesis Only NLI

Hypothesis Only NLI

Hypothesis: ***A woman is sleeping***

Hypothesis Only NLI

Premise: $\varnothing$

Hypothesis: ***A woman is sleeping***

Hypothesis Only NLI

Premise: $\varnothing$

Hypothesis: ***A woman is sleeping***

entailment neutral contradiction

Hypothesis Only NLI

Premise: ∅

Hypothesis: ***A woman is sleeping***

entailment neutral contradiction

Why is that a “contradiction”?

Why is that a “contradiction”?

Can a model pick up on this?

Why is that a “contradiction”?

Can a model pick up on this?

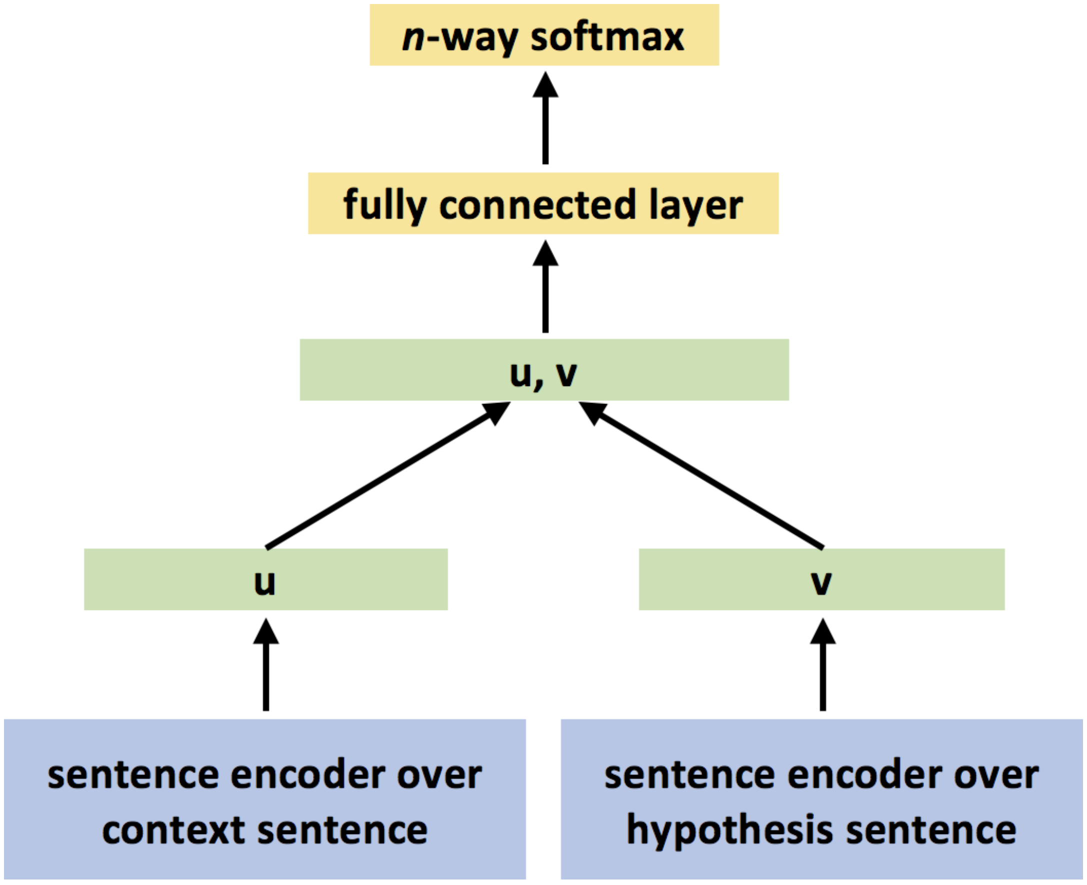
What does this say about NLI?

Do NLI datasets contain statistical irregularities that allow hypothesis only models to outperform each dataset's specific prior?

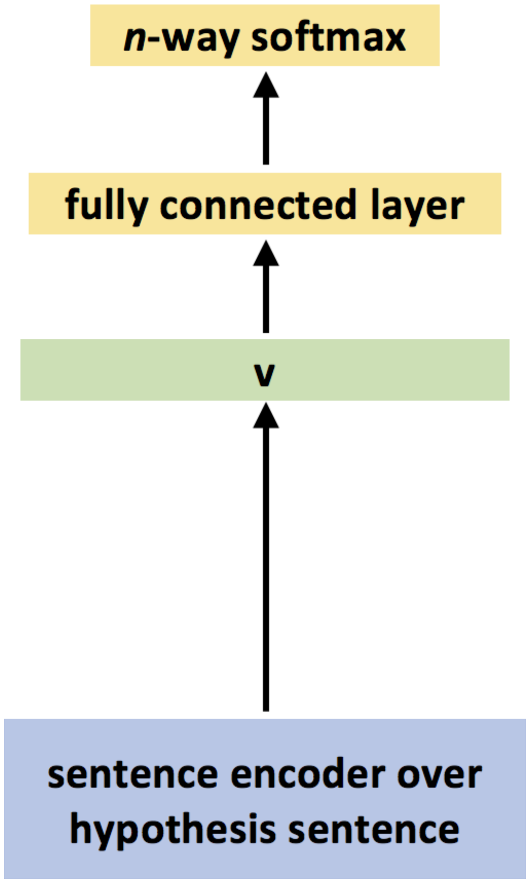
Outline

- ~~Introduction~~
- Hypothesis Only Model
- Data under investigation
- Experiments & Results

Typical NLI Model

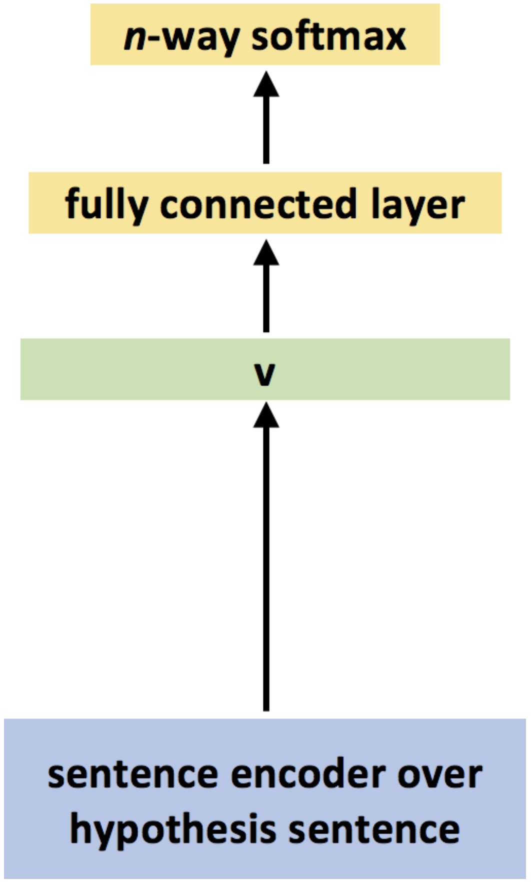


Hypothesis Only Model



Hypothesis Only Model

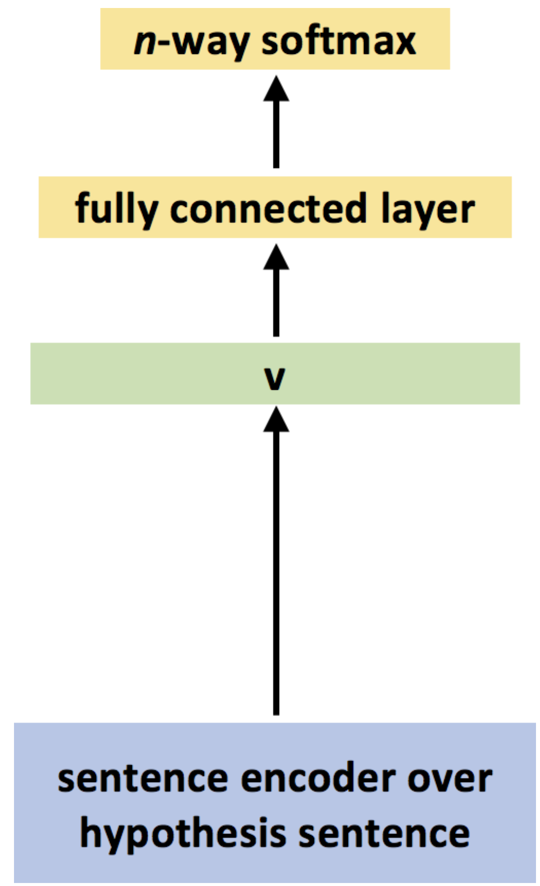
Goal:
Representative of
common NLI research



Hypothesis Only Model

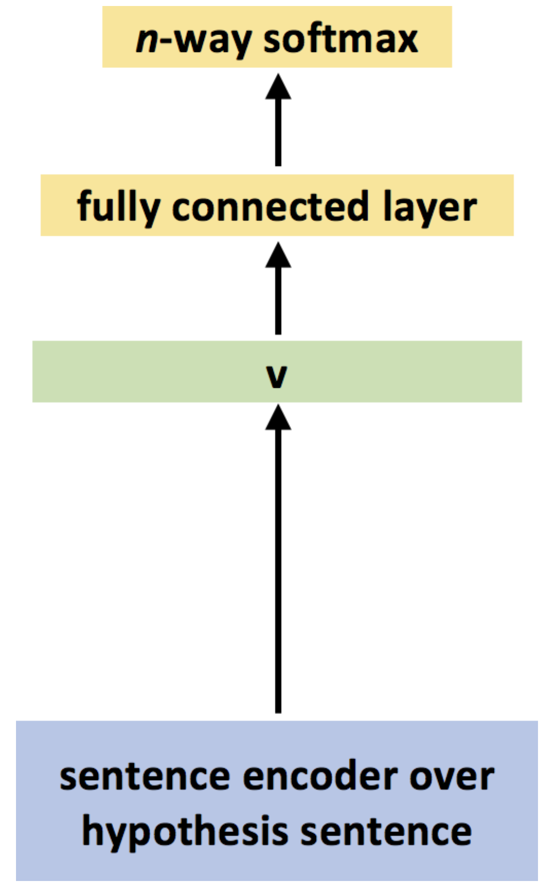
Goal:
Representative of common NLI research

No modeling contribution



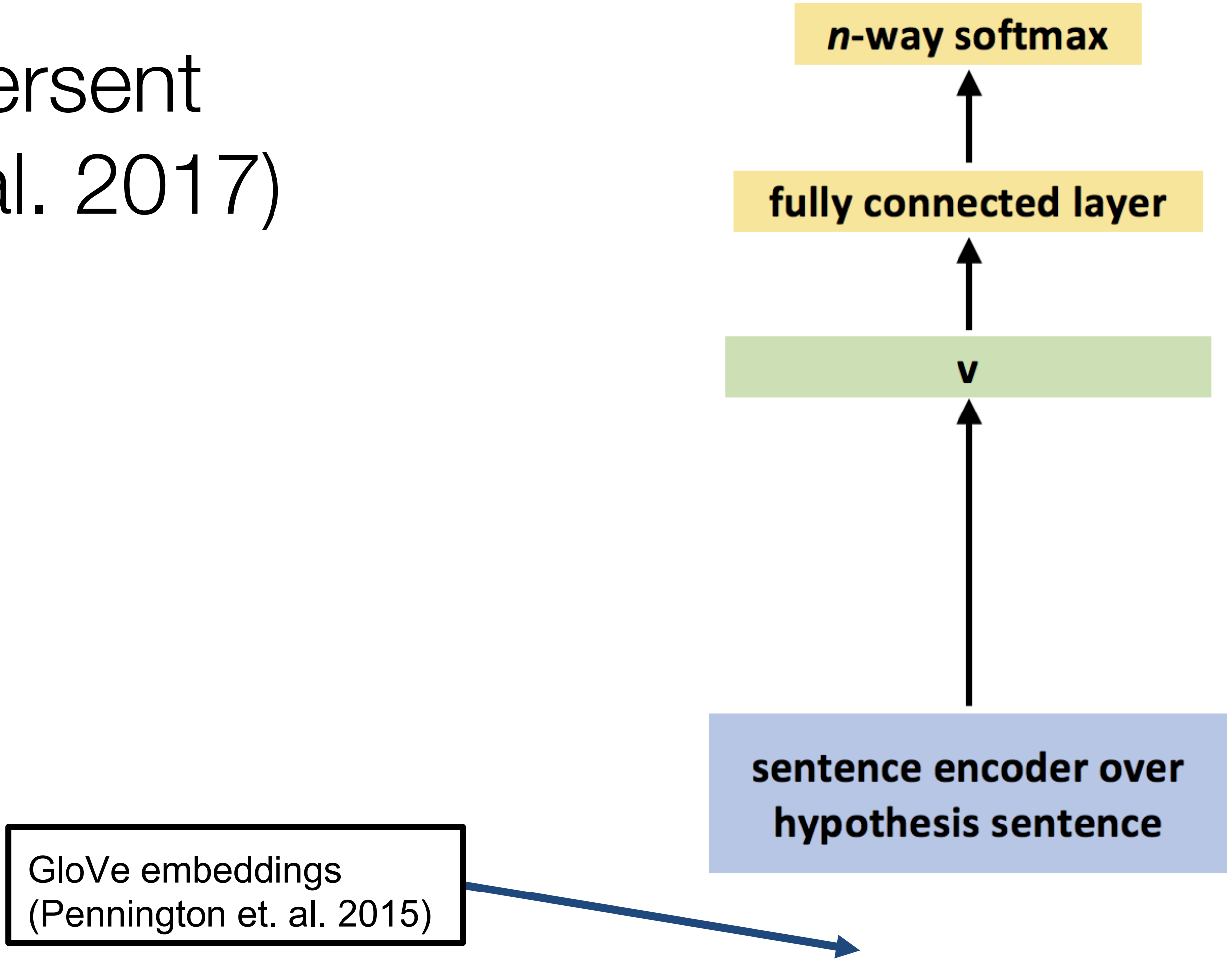
Hypothesis Only Model

Modified Infersent
(Conneau et. al. 2017)



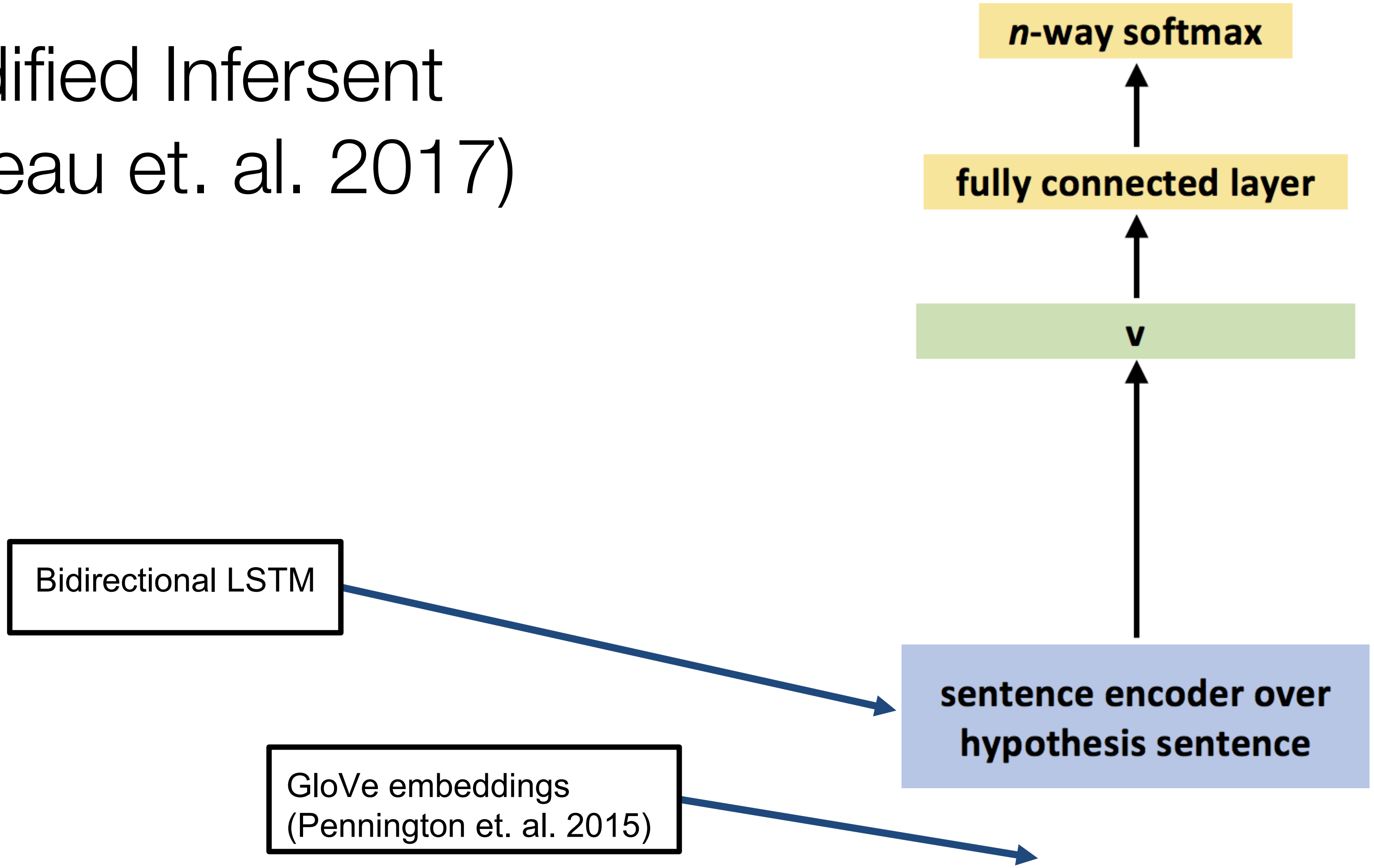
Hypothesis Only Model

Modified Infersent
(Conneau et. al. 2017)



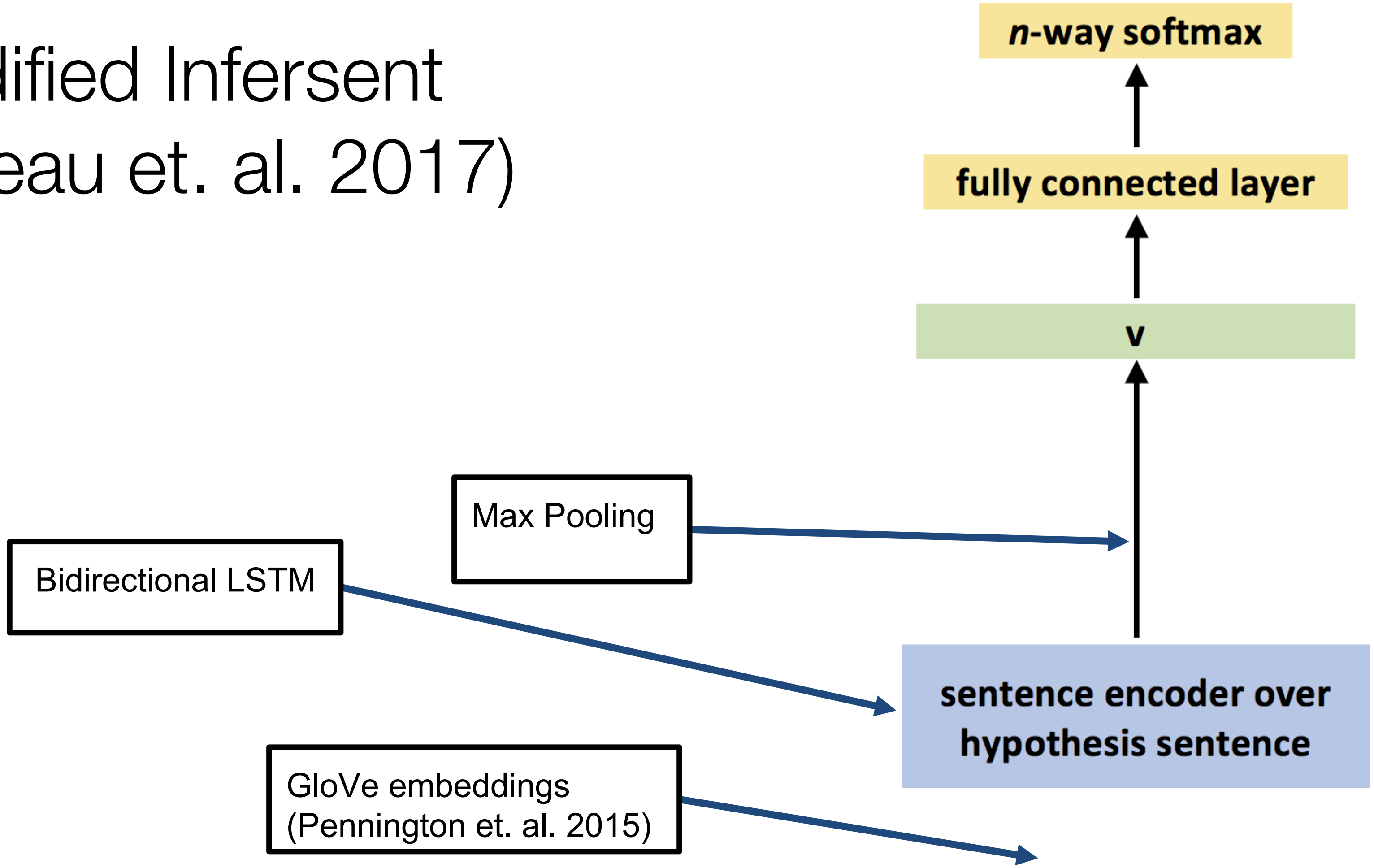
Hypothesis Only Model

Modified Infersent
(Conneau et. al. 2017)



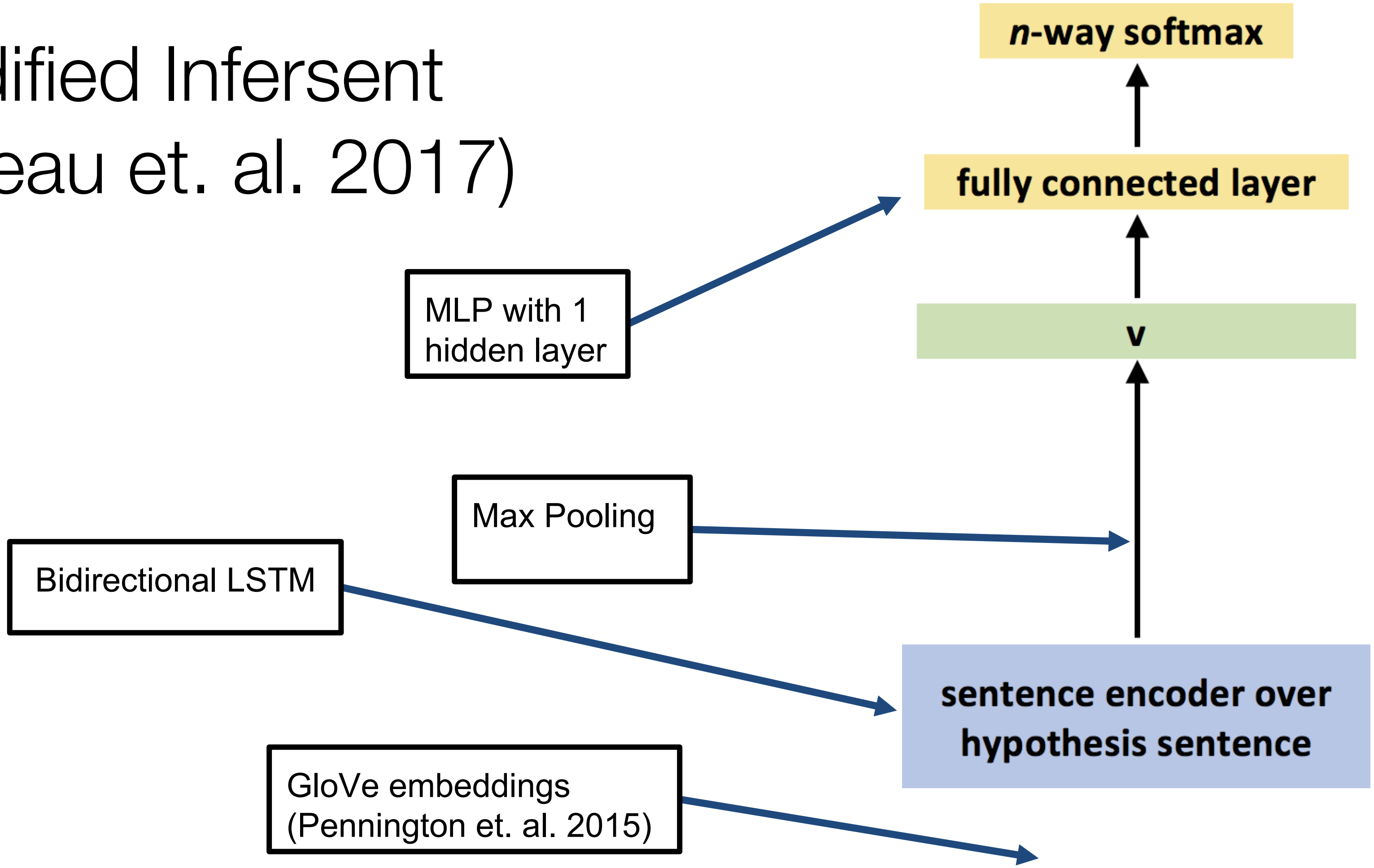
Hypothesis Only Model

Modified Infersent
(Conneau et. al. 2017)



Hypothesis Only Model

Modified Infersent
(Conneau et. al. 2017)



Outline

- ~~Introduction~~
- ~~Hypothesis Only Model~~
- Data under investigation
- Experiments & Results

Human elicited

Human is:

Human elicited

Human is:

1. shown context (premise)

Human elicited

Human is:

1. shown context (premise)
2. generates hypothesis for each label:

   entailed, neutral, contradiction

Human elicited

Human is:

1. shown context (premise)
2. generates hypothesis for each label:

   entailed, neutral, contradiction

Used in SNLI & Multi-NLI creation

Human elicited - Example

Premise: ***A woman is reading with a child***

entailment neutral contradiction

Human elicited - Example

Premise: ***A woman is reading with a child***

~~entailment~~ ~~neutral~~ contradiction

Human elicited - Example

Premise: ***A woman is reading with a child***

Hypothesis: ***A woman is sleeping***

~~entailment~~ ~~neutral~~ contradiction

Human judged

Human is:

Human judged

Human is:

1. shown context and hypothesis pair

Human judged

Human is:

1. shown context and hypothesis pair
2. assigns a label to the pair

Human judged

Human is:

1. shown context and hypothesis pair
2. assigns a label to the pair

Used in:

SICK (Marelli et. al. 2014), Add-1 (Pavlick et. al. 2016), MPE (Lai et. al. 2017), JOCI (Zhang et. al. 2017), SciTail (Khot et. al. 2018)

Recast

Recast

Minimize human annotation involvement

Recast

Minimize human annotation involvement

Annotations from existing NLU datasets recast as NLI

Recast

Minimize human annotation involvement

Annotations from existing NLU datasets recast as NLI

White et. al. (2017) recast:

- SPR (Reisinger et. al. 2016)
- FN+ (Pavlick et. al. 2015)
- DPR (Rahman & Ng 2012)

JOHNS HOPKINS UNIVERSITY

Recast Semantic Proto-Roles

Recast Semantic Proto-Roles

Premise: ***He blames imports***

Recast Semantic Proto-Roles

Premise: ***He blames imports***

Recast Semantic Proto-Roles

Premise: ***He blames imports***

sentient:
volitional:
existed after:

Recast Semantic Proto-Roles

Premise: ***He blames imports***

sentient: **X**
volitional:
existed after:

Recast Semantic Proto-Roles

Premise: ***He blames imports***

sentient: **X**
volitional: **X**
existed after:

Recast Semantic Proto-Roles

Premise: ***He blames imports***

sentient: **X**
volitional: **X**
existed after: ✓

Recast Semantic Proto-Roles

Premise: ***He blames imports***

sentient: **X**

volitional: **X**

existed after: ✓

Recast Semantic Proto-Roles

Premise: ***He blames imports***

Hypothesis: ***Imports were sentient***

entailed   not-entailed

3 Types of NLI datasets

Human Elicited

Human Judged

Recast

JOHNS HOPKINS
UNIVERSITY

Outline

- ~~Introduction~~
- ~~Hypothesis Only Model~~
- ~~Data under investigation~~
- Experiments & Results

Experimental Setup

Experimental Setup

Train a hypothesis only model on each dataset

Experimental Setup

Train a hypothesis only model on each dataset

Test the model on each specific dataset

JOHNS HOPKINS UNIVERSITY

Experimental Setup

Train a hypothesis only model on each dataset

Test the model on each specific dataset

Compare hypothesis only model to majority baseline

Results across 10 datasets

Results across 10 datasets

	DEV				TEST					
Dataset	Hyp-Only	MAJ	REL-ABS	REL-%	Hyp-Only	MAJ	REL-ABS	REL-%	Baseline	SOTA
					Recast					
DPR	50.21	50.21	0.00	0.00	49.95	49.95	0.00	0.00	49.5	49.5
SPR	86.21	65.27	+20.94	+32.08	86.57	65.44	+21.13	+32.29	80.6	80.6
FN+	62.43	56.79	+5.64	+9.31	61.11	57.48	+3.63	+6.32	80.5	80.5
					Human Judged					
ADD-1	75.10	75.10	0.00	0.00	85.27	85.27	0.00	0.00	92.2	92.2
SciTail	66.56	50.38	+16.18	+32.12	66.56	60.04	+6.52	+10.86	70.6	77.3
SICK	56.76	56.76	0.00	0.00	56.87	56.87	0.00	0.00	56.87	84.6
MPE	40.20	40.20	0.00	0.00	42.40	42.40	0.00	0.00	41.7	56.3
JOCI	61.64	57.74	+3.90	+6.75	62.61	57.26	+5.35	+9.34	–	–
					Human Elicited					
SNLI	69.17	33.82	+35.35	+104.52	69.00	34.28	+34.72	+101.28	78.2	89.3
MNLI-1	55.52	35.45	+20.07	+56.61	–	35.6	– –		72.3	80.60
MNLI-2	55.18	35.22	+19.96	+56.67	–	36.5	–	–	72.1	83.21

Results across 10 datasets

	DEV				TEST					
Dataset	Hyp-Only	MAJ	REL-ABS	REL-%	Hyp-Only	MAJ	REL-ABS	REL-%	Baseline	SOTA
					Recast					
DPR	50.21	50.21	0.00	0.00	49.95	49.95	0.00	0.00	49.5	49.5
SPR	86.21	65.27	+20.94	+32.08	86.57	65.44	+21.13	+32.29	80.6	80.6
FN+	62.43	56.79	+5.64	+9.31	61.11	57.48	+3.63	+6.32	80.5	80.5
					Human Judged					
ADD-1	75.10	75.10	0.00	0.00	85.27	85.27	0.00	0.00	92.2	92.2
SciTail	66.56	50.38	+16.18	+32.12	66.56	60.04	+6.52	+10.86	70.6	77.3
SICK	56.76	56.76	0.00	0.00	56.87	56.87	0.00	0.00	56.87	84.6
MPE	40.20	40.20	0.00	0.00	42.40	42.40	0.00	0.00	41.7	56.3
JOCI	61.64	57.74	+3.90	+6.75	62.61	57.26	+5.35	+9.34	–	–
					Human Elicited					
SNLI	69.17	33.82	+35.35	+104.52	69.00	34.28	+34.72	+101.28	78.2	89.3
MNLI-1	55.52	35.45	+20.07	+56.61	–	35.6	– –		72.3	80.60
MNLI-2	55.18	35.22	+19.96	+56.67	–	36.5	–	–	72.1	83.21

Results across 10 datasets

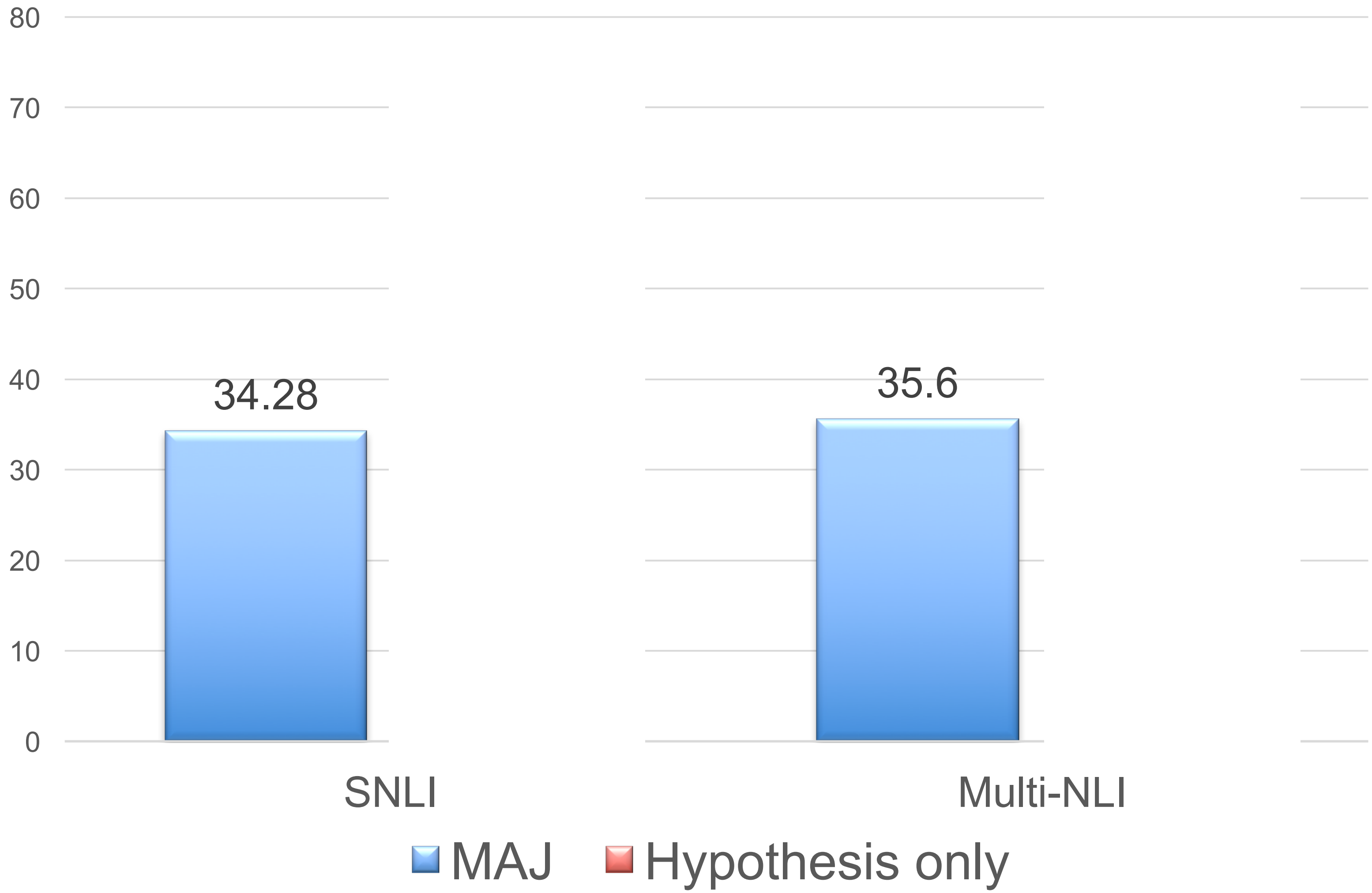
	DEV				TEST					
Dataset	Hyp-Only	MAJ	REL-ABS	REL-%	Hyp-Only	MAJ	REL-ABS	REL-%	Baseline	SOTA
Recast										
DPR	50.21	50.21	0.00	0.00	49.95	49.95	0.00	0.00	49.5	49.5
SPR	86.21	65.27	+20.94	+32.08	86.57	65.44	+21.13	+32.29	80.6	80.6
FN+	62.43	56.79	+5.64	+9.31	61.11	57.48	+3.63	+6.32	80.5	80.5
Human Judged										
ADD-1	75.10	75.10	0.00	0.00	85.27	85.27	0.00	0.00	92.2	92.2
SciTail	66.56	50.38	+16.18	+32.12	66.56	60.04	+6.52	+10.86	70.6	77.3
SICK	56.76	56.76	0.00	0.00	56.87	56.87	0.00	0.00	56.87	84.6
MPE	40.20	40.20	0.00	0.00	42.40	42.40	0.00	0.00	41.7	56.3
JOCI	61.64	57.74	+3.90	+6.75	62.61	57.26	+5.35	+9.34	–	–
Human Elicited										
SNLI	69.17	33.82	+35.35	+104.52	69.00	34.28	+34.72	+101.28	78.2	89.3
MNLI-1	55.52	35.45	+20.07	+56.61	[illegible]	35.6	[illegible]	[illegible]	72.3	80.60
MNLI-2	55.18	35.22	+19.96	+56.67	–	36.5	–	–	72.1	83.21

Statistical Irregularities
or
Background Knowledge?

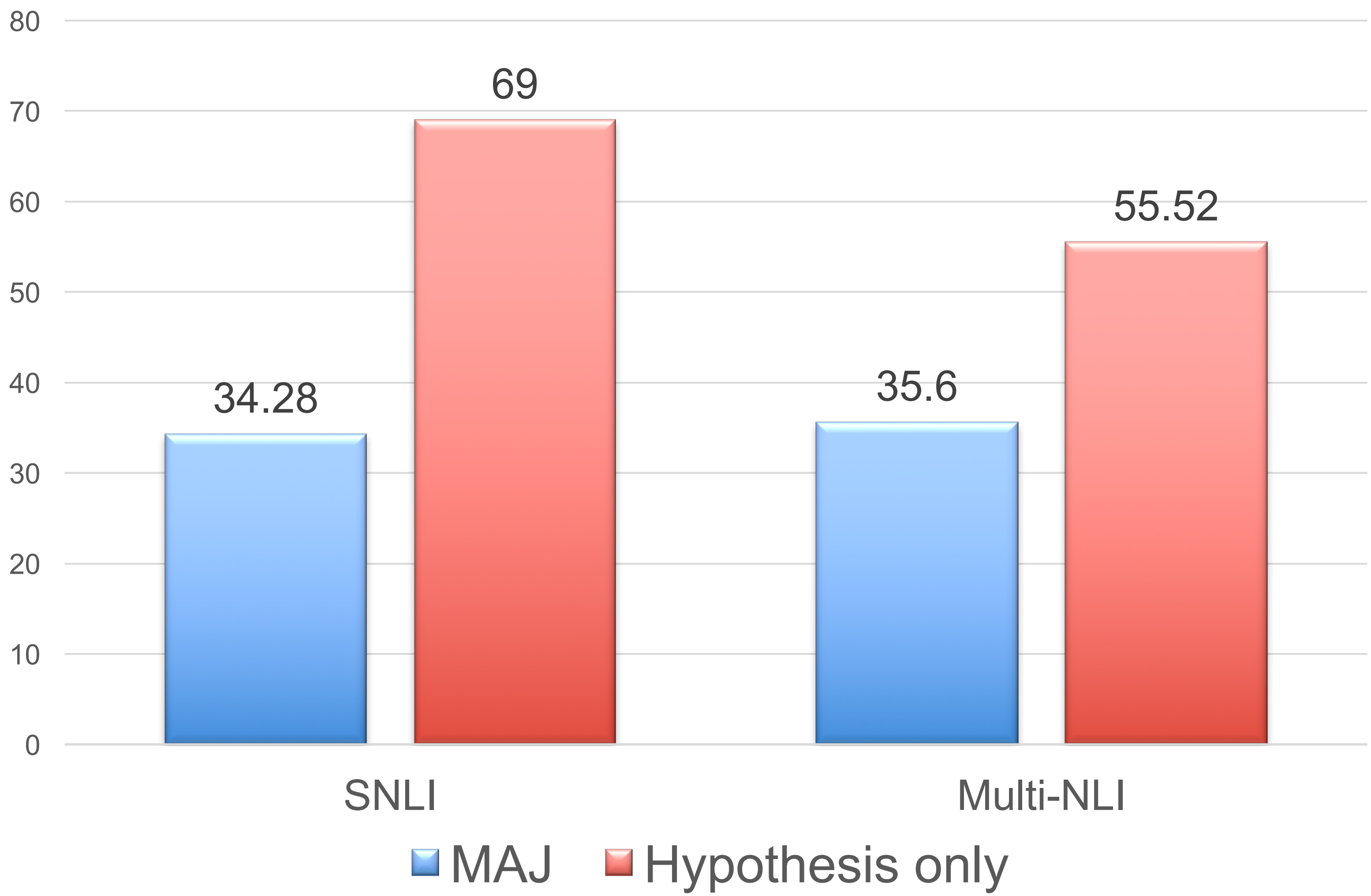
JOHNS HOPKINS
UNIVERSITY

Human Elicited Results

Human Elicited Results



Human Elicited Results



JOHNS HOPKINS UNIVERSITY

Origin of SNLI

Origin of SNLI

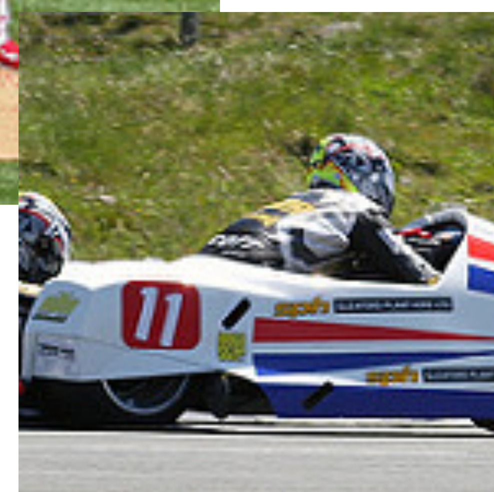
flickr

(Young et. al. 2014)

Origin of SNLI

flickr

(Young et. al. 2014)



A woman is sleeping

Premises:

Hypothesis: A woman is sleeping

Premises:

A woman sings a song while playing piano

Hypothesis: A woman is sleeping

Premises:

This woman is laughing at her baby shower

Hypothesis: A woman is sleeping

Premises:

A woman with glasses is playing jenga

Hypothesis: A woman is sleeping

Why is she sleeping?

Studies in eliciting norming data are prone to **repeated responses across subjects**

Elicitation Bias

Descriptions of "dog":

-- McRae et al. (2005)

Elicitation Bias

Descriptions of “dog”:

- barks

-- McRae et al. (2005)

JOHNS HOPKINS UNIVERSITY

Elicitation Bias

Descriptions of “dog”:

- barks
- has a tail

-- McRae et al. (2005)

JOHNS HOPKINS UNIVERSITY

Elicitation Bias

Descriptions of “dog”:

- barks
- has a tail
- larger than a tulip

-- McRae et al. (2005)

JOHNS HOPKINS UNIVERSITY

Elicitation Bias

Descriptions of “dog”:

- barks
- has a tail
- larger than a tulip
- moves faster than an infant

-- McRae et al. (2005)

JOHNS HOPKINS UNIVERSITY

Elicitation Bias

*"Features such as **is larger than a tulip** or **moves faster than an infant**, although logically possible, do not occur in human responses … people are capable of **verifying** that a **dog is larger than a pencil**."*

-- McRae et al. (2005)

JOHNS HOPKINS UNIVERSITY

Studies in eliciting norming data are prone to **repeated responses across subjects**

(see discussion in §2 of Zhang et. al. (2017))

JOHNS HOPKINS UNIVERSITY

Inferring labels from single words

“Give away” words

$$p(l|w) = \frac{count(w,l)}{count(w)}$$

“Give away” words

$$p(l|w) = \frac{count(w,l)}{count(w)}$$

$$p(l|w) > \alpha$$

Words correlated with contradictions

Word	$p(l\|w)$	Frequency

Words correlated with contradictions

Word	$p(l \vert w)$	Frequency
sleeping	0.88	108
asleep	0.91	43
sleeps	0.95	20

Words correlated with contradictions

Word	$p(l\|w)$	Frequency
Nobody	1.00	52
alone	0.90	50
no	0.84	31
empty	0.93	28

JOHNS HOPKINS UNIVERSITY

Words correlated with contradictions

Word	$p(l\|w)$	Frequency
driving	0.81	53
eats	0.83	24

Recast NLI

	DEV				TEST					
Dataset	Hyp-Only	MAJ	REL-ABS	REL-%	Hyp-Only	MAJ	REL-ABS	REL-%	Baseline	SOTA
				Recast						
DPR	50.21	50.21	0.00	0.00	49.95	49.95	0.00	0.00	49.5	49.5
SPR	86.21	65.27	+20.94	+32.08	86.57	65.44	+21.13	+32.29	80.6	80.6
FN+	62.43	56.79	+5.64	+9.31	61.11	57.48	+3.63	+6.32	80.5	80.5
				Human Judged						
ADD-1	75.10	75.10	0.00	0.00	85.27	85.27	0.00	0.00	92.2	92.2
SciTail	66.56	50.38	+16.18	+32.12	66.56	60.04	+6.52	+10.86	70.6	77.3
SICK	56.76	56.76	0.00	0.00	56.87	56.87	0.00	0.00	56.87	84.6
MPE	40.20	40.20	0.00	0.00	42.40	42.40	0.00	0.00	41.7	56.3
JOCI	61.64	57.74	+3.90	+6.75	62.61	57.26	+5.35	+9.34	–	–
				Human Elicited						
SNLI	69.17	33.82	+35.35	+104.52	69.00	34.28	+34.72	+101.28	78.2	89.3
MNLI-1	55.52	35.45	+20.07	+56.61	–	35.6	– –		72.3	80.60
MNLI-2	55.18	35.22	+19.96	+56.67	–	36.5	–	–	72.1	83.21

Semantic Proto-Roles

Dowty (1990)'s fine-grained version of thematic roles

Proto-Agent & *Proto-Patient* properties

Dataset released by Reisinger et. al. (2015) & White et. al. (2016)

Recast Semantic Proto-Roles

Premise: ***He blames imports***

Hypothesis: ***Imports were sentient***

entailed　　　　　　not-entailed

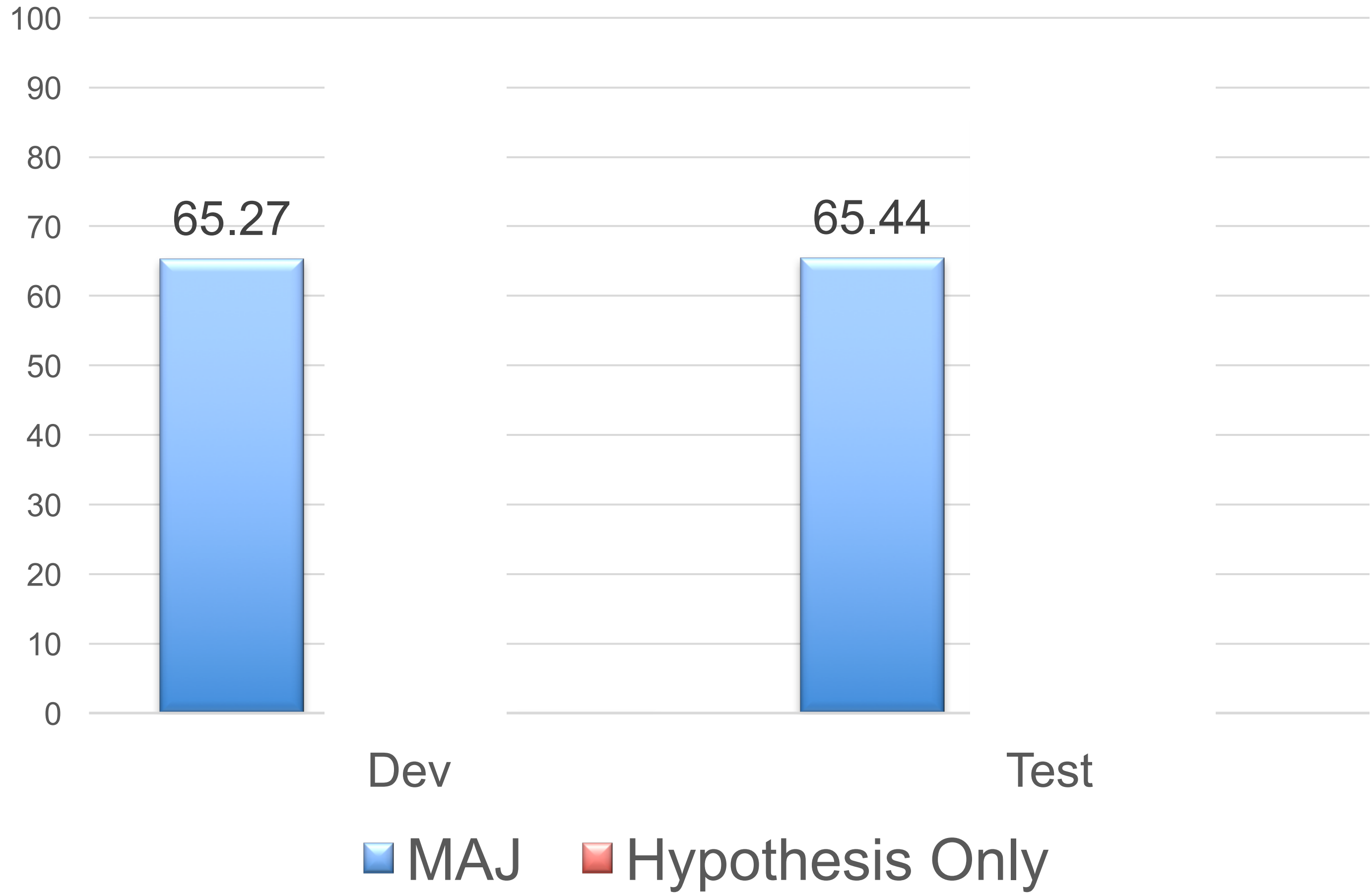
Recast Semantic Proto-Roles

Premise: ***He blames imports***

Hypothesis: ***Imports existed after the blaming***

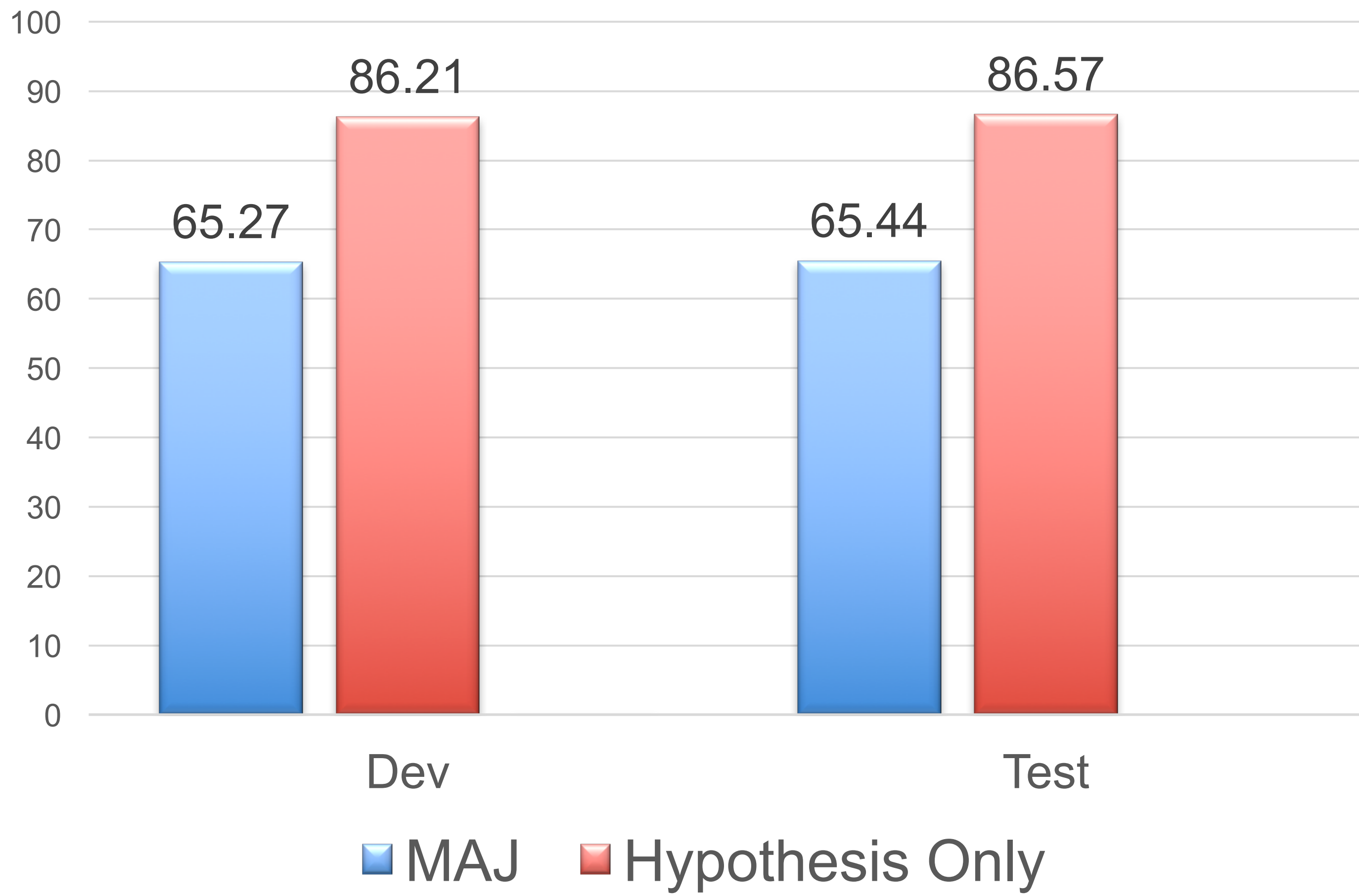
entailed   not-entailed

Hypothesis Only SPR Results

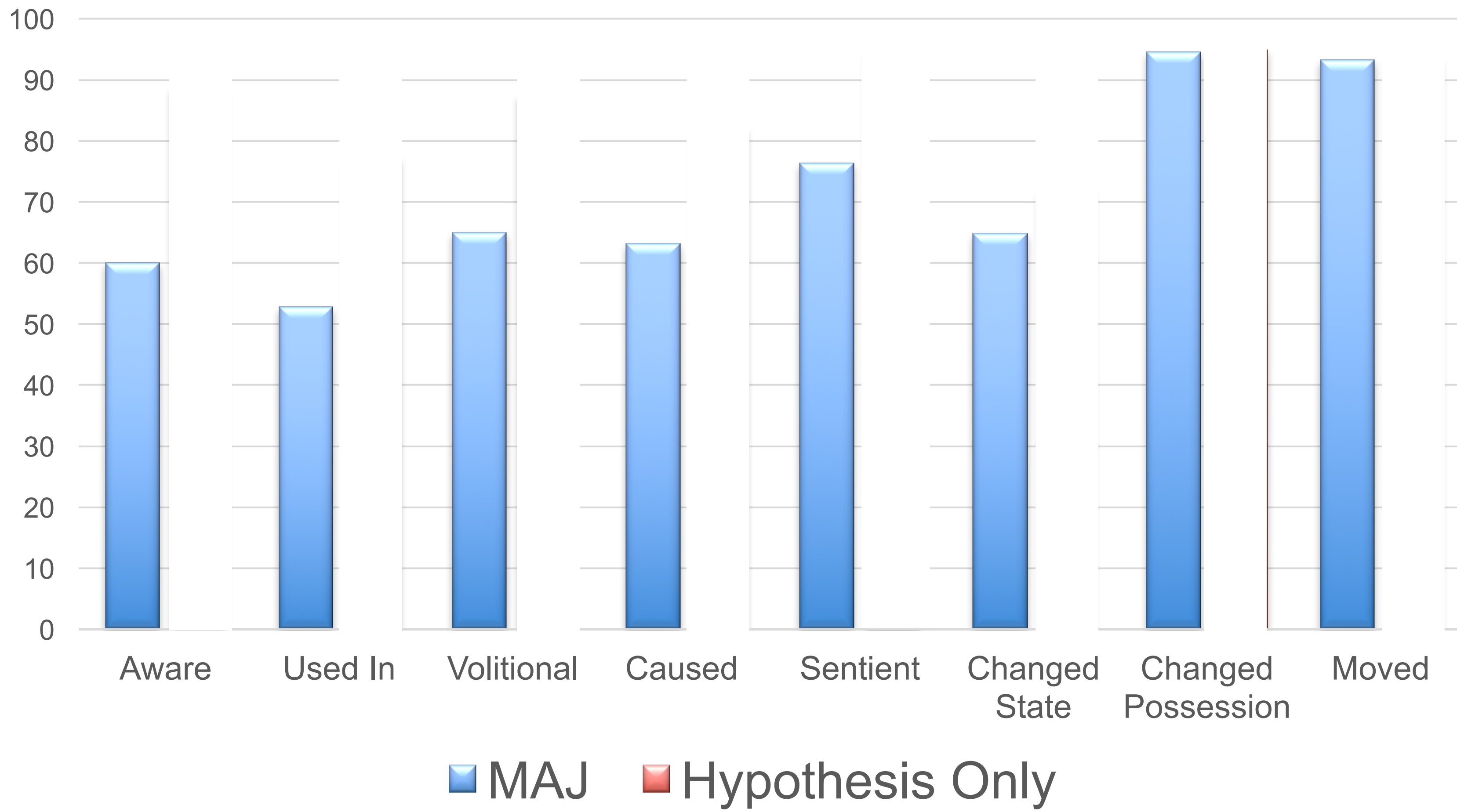


JOHNS HOPKINS UNIVERSITY

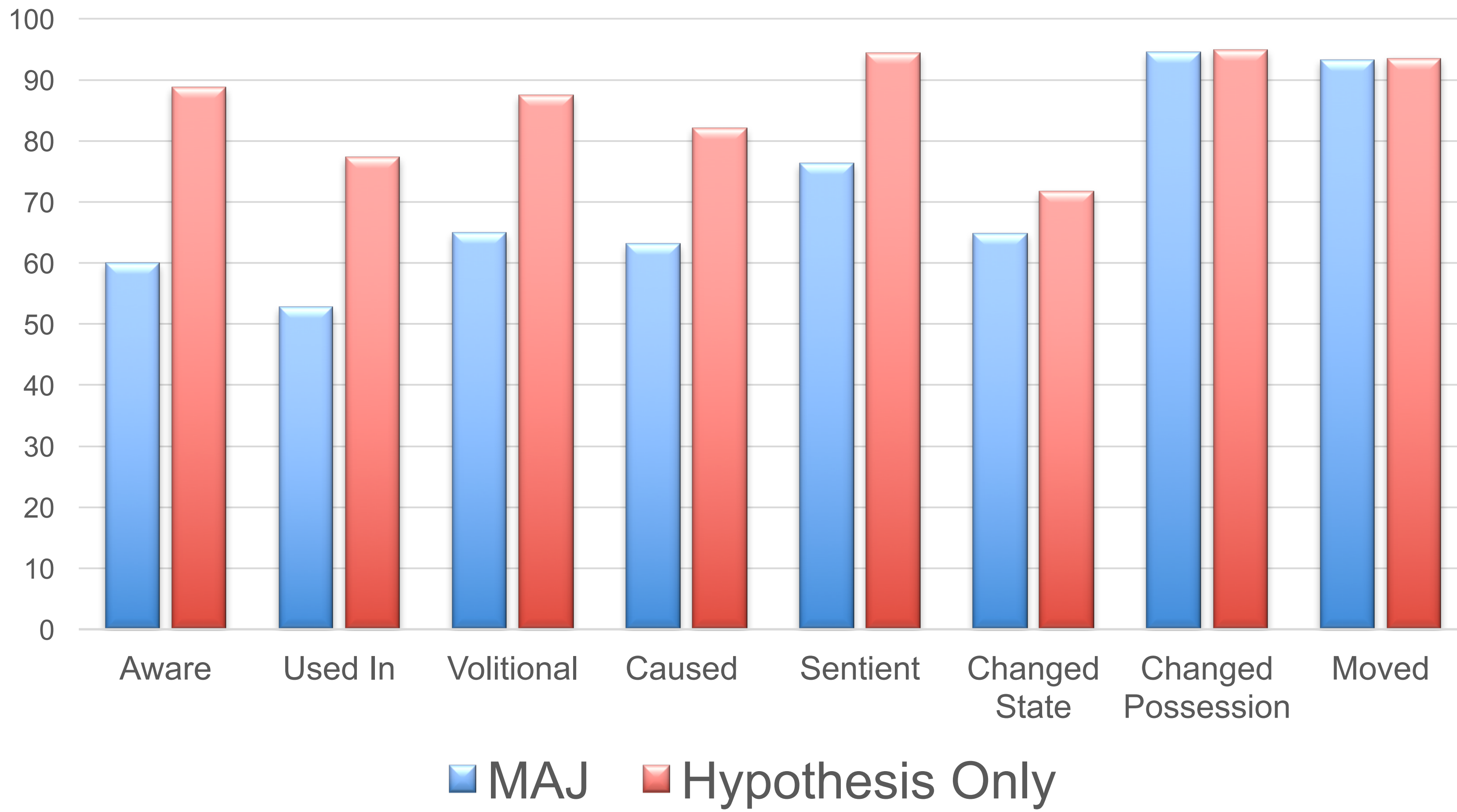
Hypothesis Only SPR Results



SPR Properties



SPR Properties



Is this surprising?

Is this surprising?

Inherent Likelihood of SPR properties

Is this surprising?

Inherent Likelihood of SPR properties

Hypotheses:

Is this surprising?

Inherent Likelihood of SPR properties

Hypotheses:

- Experts were sentient

JOHNS HOPKINS UNIVERSITY

Is this surprising?

Inherent Likelihood of SPR properties

Hypotheses:

- Experts were sentient
- Mr. Falls was sentient

Is this surprising?

Inherent Likelihood of SPR properties

Hypotheses:

- Experts were sentient
- Mr. Falls was sentient
- The campaign was sentient

Is this surprising?

Inherent Likelihood of SPR properties

Hypotheses:

- Experts were sentient
- Mr. Falls was sentient
- The campaign was sentient
    - probably not

Is this surprising? *No*

Inherent Likelihood of SPR properties

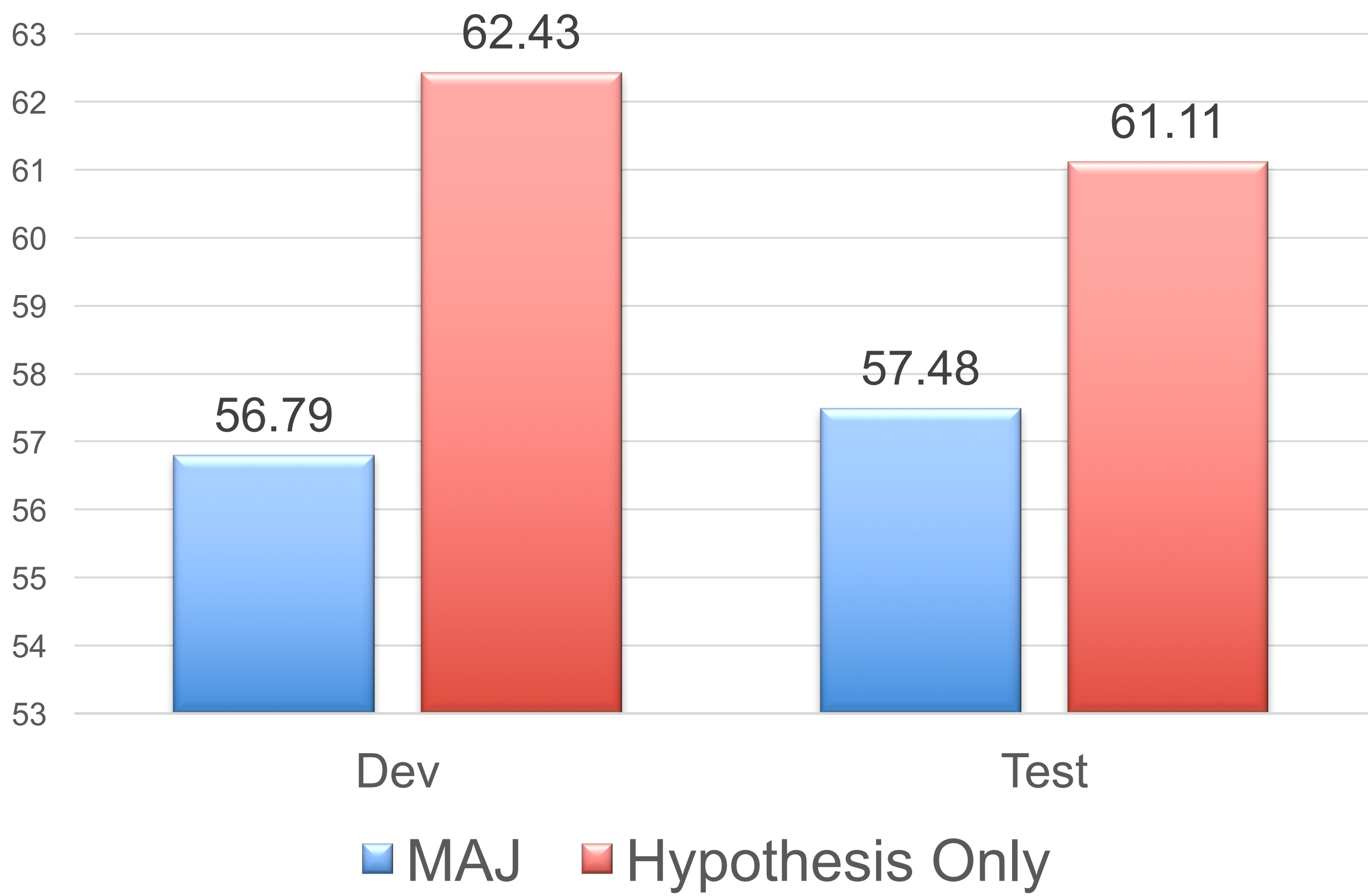
Hypotheses:

- Experts were sentient
- Mr. Falls was sentient
- The campaign was sentient
  - probably not

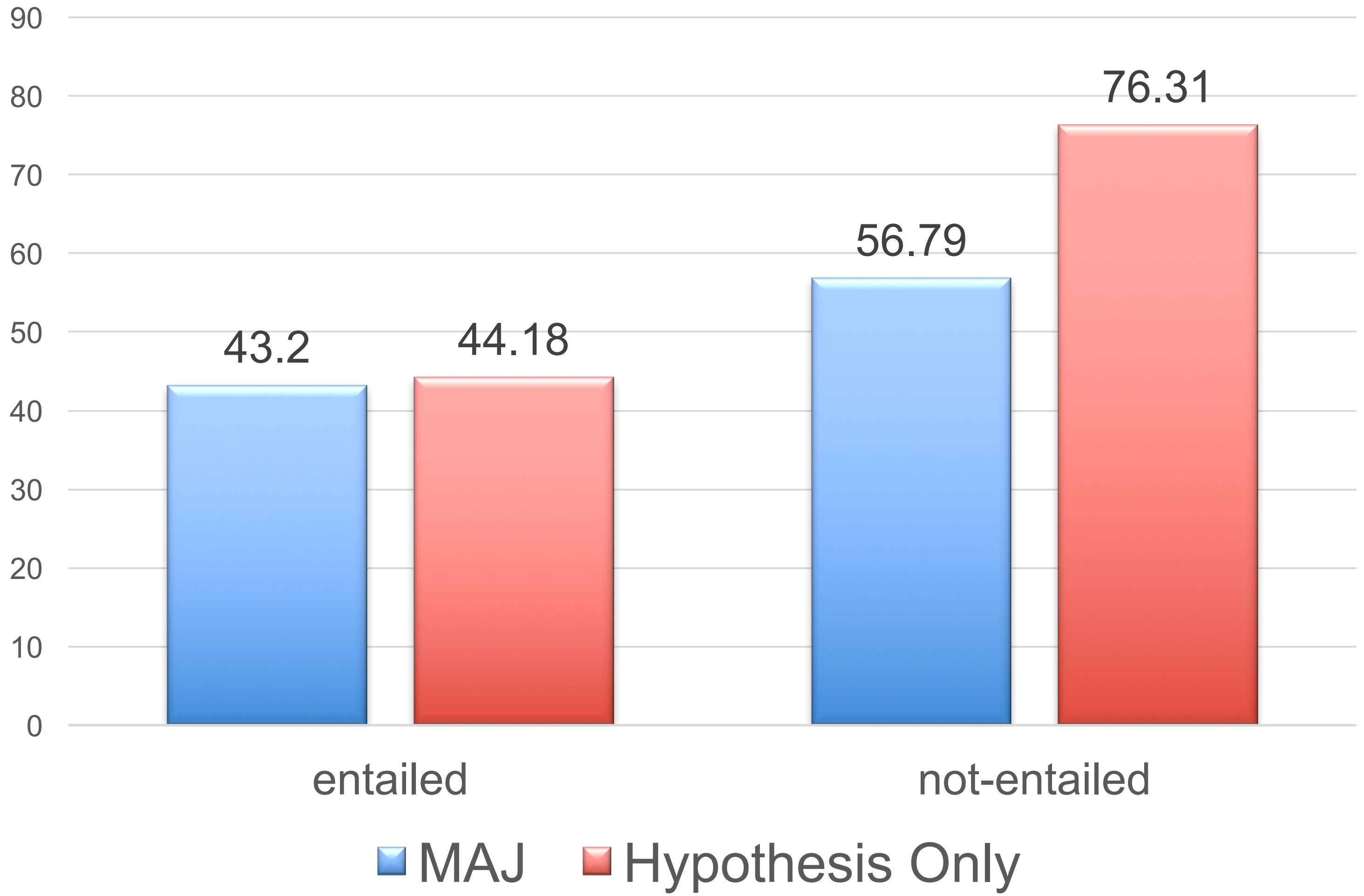
JOHNS HOPKINS UNIVERSITY

Hypothesis Only Recast FN+ Results

Hypothesis Only Recast FN+ Results



Recast FN+ Results by Label



Recast FN+
(Paraphrastic Inference)

JOHNS HOPKINS UNIVERSITY

Recast FN+ (Paraphrastic Inference)

Swap single tokens based on PPDB

Recast FN+ (Paraphrastic Inference)

Swap single tokens based on PPDB

entailed: high scoring paraphrase

Entailed FN+ Example

Premise: *Jerusalem fell to the Ottomans in 1517, remaining under their* ***control*** *for 400 years*

control -> supervision

Entailed FN+ Example

Premise: *Jerusalem fell to the Ottomans in 1517, remaining under their* ***<u>control</u>*** *for 400 years*

Hypothesis: *Jerusalem fell to the Ottomans in 1517, remaining under their* ***<u>supervision</u>*** *for 400 years*

Recast FN+ (Paraphrastic Inference)

Swap single tokens based on PPDB

not-entailed: low scoring paraphrase

Not-Entailed FN+ Example

Premise: *Jerusalem fell to the Ottomans in 1517, remaining under their **<u>control</u>** for 400 years*

control** -> **regulate

Not-Entailed FN+ Example

Premise: *Jerusalem fell to the Ottomans in 1517, remaining under their* ***control*** *for 400 years*

Hypothesis: *Jerusalem fell to the Ottomans in 1517, remaining under their* ***regulate*** *for 400 years*

FN+

Statistical Irregularities
or
Background Knowledge

JOHNS HOPKINS UNIVERSITY

FN+ Hypotheses

Entailed Hypothesis: *Jerusalem fell to the Ottomans in 1517, remaining under their* ***supervision*** *for 400 years*

Not-Entailed Hypothesis: *Jerusalem fell to the Ottomans in 1517, remaining under their* ***regulate*** *for 400 years*

FN+

Statistical Irregularities

or

Background Knowledge

Purpose of NLI (as NLP Task)

Purpose of NLI (as NLP Task)

Evaluate a model for NLU

Purpose of NLI (as NLP Task)

Evaluate a model for NLU

FraCas (Cooper et. al. 1996)

RTE (Glickman 2006, *i.a.*)

Purpose of NLI (as NLP Task)

Evaluate a model for NLU

FraCas (Cooper et. al. 1996)

RTE (Glickman 2006, *i.a.*)

Train a model for NLU

Purpose of NLI (as NLP Task)

Evaluate a model for NLU

FraCas (Cooper et. al. 1996)

RTE (Glickman 2006, *i.a.*)

Train a model for NLU

SNLI (Bowman et. al. 2015)

Multi-NLI (Williams et. al. 2018)

Prior Non-archival Work

Sitzmann, Marek, Keselman (Stanford Course Project 2016)

Concurrent Work

Masatoshi Tsuchiya (LREC2018)

Gururangan, Swayamdipta, Levy, Schwartz, Bowman, and Smith (NAACL 2018)

Concurrent Work

"Hypothesis sentences of the SNLI corpus are composed by human workers, but all sentences of the SICK corpus are derived from original sentences using hand-crafted rules. We think that ***this difference may be a cause of the hidden bias revealed by this paper****"*

Tsuchiya (LREC2018)

Concurrent Work

"We show that, in a significant portion of such data, ***this protocol leaves clues that make it possible to identify the label by looking only at the hypothesis****, without observing the premise"*

Gururangan et. al. (NAACL 2018)

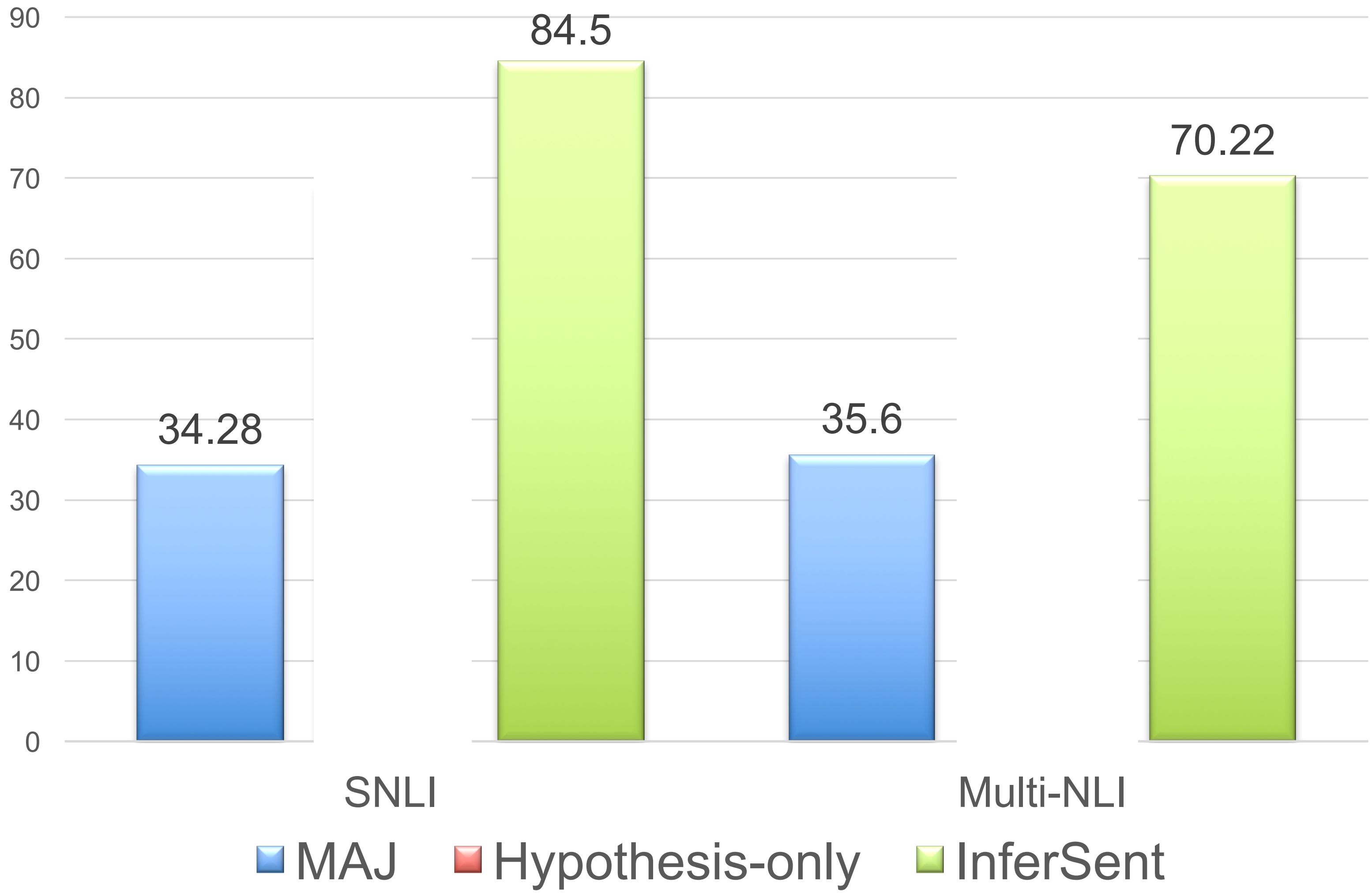
Summary

Human elicitation has biases but might not be statistical irregularities

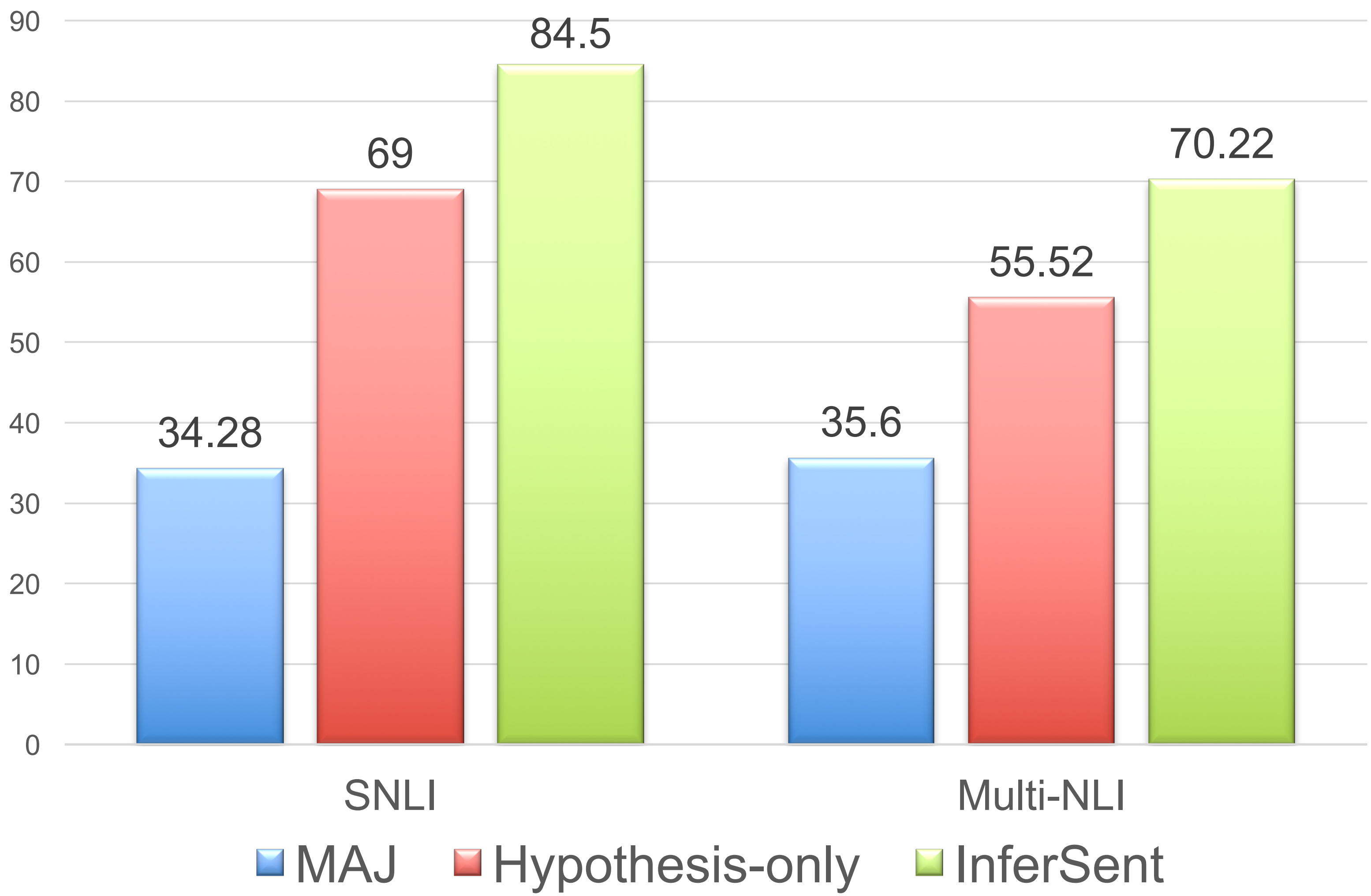
Recasting methods may create statistical irregularities

Compare NLI models with corresponding hypothesis only version

InferSent


90
80
70
60
50
40
30
20
10
0
34.28
84.5
35.6
70.22
SNLI
Multi-NLI
MAJ
Hypothesis-only
InferSent

InferSent



90
80
70
60
50
40
30
20
10
0
34.28
69
84.5
35.6
55.52
70.22
SNLI
Multi-NLI
MAJ
Hypothesis-only
InferSent

Thank you

Jason Naradowsky

Aparajita Haldar

Rachel Rudinger

Benjamin Van Durme

JOHNS HOPKINS UNIVERSITY