Investigating the Generalization Ability of Neural Models through Monotonicity Reasoning

Hitomi Yanaka
RIKEN
http://hitomiyanaka.strikingly.com/
NALOMA@WeSSLLI2020 July 15, 2020

Natural Language Inference (NLI)
aka, Recognizing Textual Entailment [Dagan+, 2013]

Does a premise P entail a hypothesis H?

P: There is no white dog leaning on the fence.

H1: There is no white multese dog leaning on the fence. *Entailment*

H2: There is no dog leaning on the fence. *Non-entailment*

State-of-the-art Deep Neural Networks (DNN) for NLI

Recent progress on neural models often updates the SOTA NLI model

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5	51.2
2	ERNIE Team - Baidu	ERNIE		90.4	74.4	97.5	93.5/91.4	93.0/92.6	75.2/90.9	91.4	91.0	96.6	90.9	94.5	51.7
3	Alibaba DAMO NLP	StructBERT		90.3	75.3	97.1	93.9/91.9	93.0/92.5	74.8/91.0	90.9	90.7	96.4	90.2	94.5	49.1
4	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1
5	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
6	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	90.8	95.8	89.8	91.8	50.7
7	Huawei Noah's Ark Lab	NEZHA-Large		88.7	67.4	97.2	93.2/91.0	92.2/91.6	74.1/90.2	90.8	90.2	95.7	88.5	93.2	45.0
8	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.7	89.0	50.1
9	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3
10	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.2	89.0	48.7
11	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
12	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-

GLUE [Wang+ 2019] Leaderboard:https://gluebenchmark.com/leaderboard

State-of-the-art Deep Neural Networks (DNN) for NLI

Recent progress on neural models often updates the SOTA NLI model

Model	MultiNLI [Williams+2018]
Human baseline	92.8
T5	92.2
ALBERT+DAAF+NAS	91.6
ERNIE	91.4
ELECTRA-Large	91.3
⋮	

Rank	Name	WNLI	AX
1	PING-AN Omni-S	94.5	51.2
2	ERNIE Team - B	94.5	51.7
3	Alibaba DAMO N	94.5	49.1
4	T5 Team - Googl	94.5	53.1
5	Microsoft D365 A	94.5	50.2
6	ELECTRA Team	91.8	50.7
7	Huawei Noah's A	93.2	45.0
8	Microsoft D365 A	89.0	50.1
9	Junjie Yang	89.0	49.3
10	Facebook AI	89.0	48.7
11	Microsoft D365 A	89.0	42.8
12	GLUE Human Ba	95.9	-

GLUE [Wang+ 2019] Leaderboard:https://gluebenchmark.com/leaderboard

Generalization Concern about DNN-based NLI

SOTA DNN models fail to perform challenging inferences

… because DNN models might learn undesired biases [Gururangan+2018] and heuristics [Mccoy+2019]

Example in the challenging NLI dataset, HANS [Mccoy+2019]

P: The lawyer mentioned the actor.

H: The actor mentioned the lawyer. *Non-entailment*

Generalization Concern about DNN-based NLI

SOTA DNN models fail to perform challenging inferences

… because DNN models might learn undesired biases [Gururangan+2018] and heuristics [Mccoy+2019]

Example in the challenging NLI dataset, HANS [Mccoy+2019]

P: The lawyer mentioned the actor.

H: The actor mentioned the lawyer. *Non-entailment*

Question:

To what extent DNN models can learn the compositional generalization capacity underlying NLI?

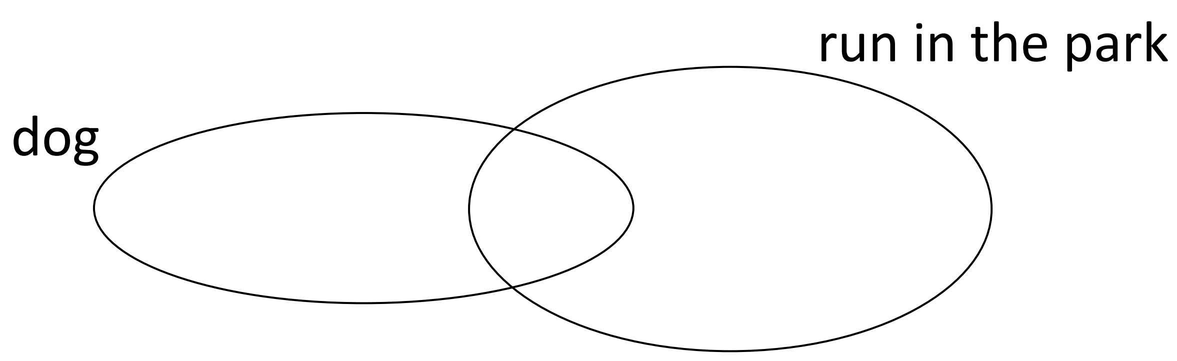
Monotonicity Reasoning [van Benthem, 1983; Icard and Moss, 2014]

- Replacements with more general (or specific) phrases license entailment

Monotonicity Reasoning [van Benthem, 1983; Icard and Moss, 2014]

- Replacements with more general (or specific) phrases license entailment
- Upward inferences: inferences from specific to general phrases

P: **Some** [dogs↑] ran in the park

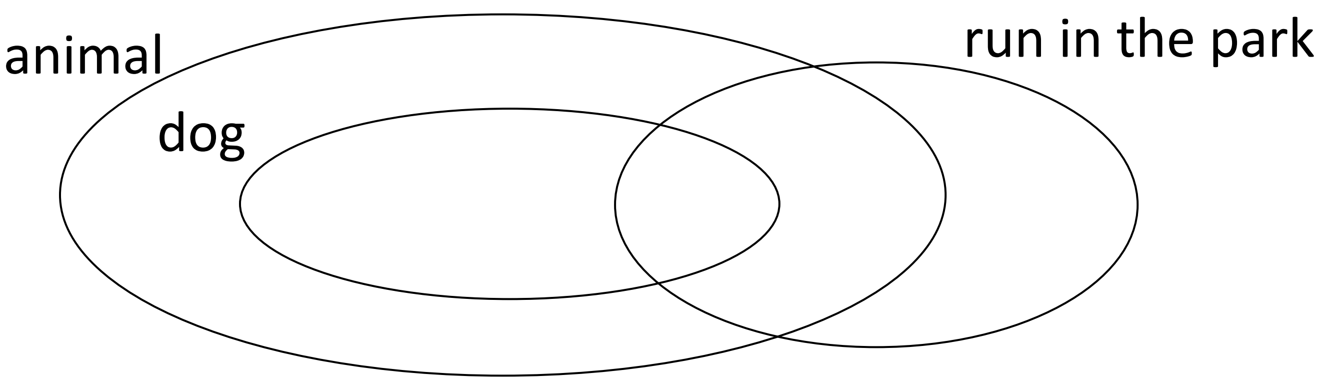


Monotonicity Reasoning [van Benthem, 1983; Icard and Moss, 2014]

- Replacements with more general (or specific) phrases license entailment
- Upward inferences: inferences from specific to general phrases

P: **Some** [dogs↑] ran in the park

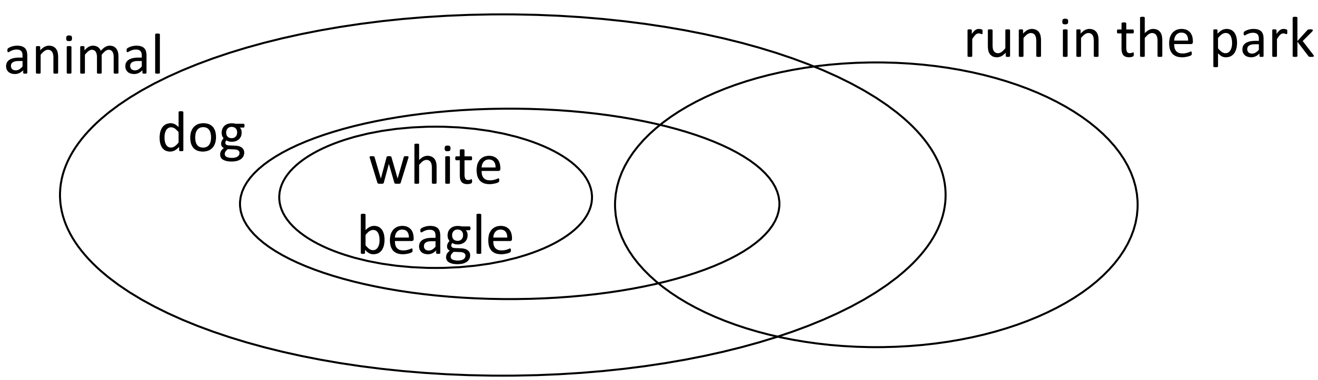
H1: **Some** [animals] ran in the park *Entailment*



Monotonicity Reasoning [van Benthem, 1983; Icard and Moss, 2014]

- Replacements with more general (or specific) phrases license entailment
- Upward inferences: inferences from specific to general phrases

P: **Some** [dogs↑] ran in the park
H1: **Some** [animals] ran in the park *Entailment*
H2: **Some** [white beagles] ran in the park *Non-Entailment*



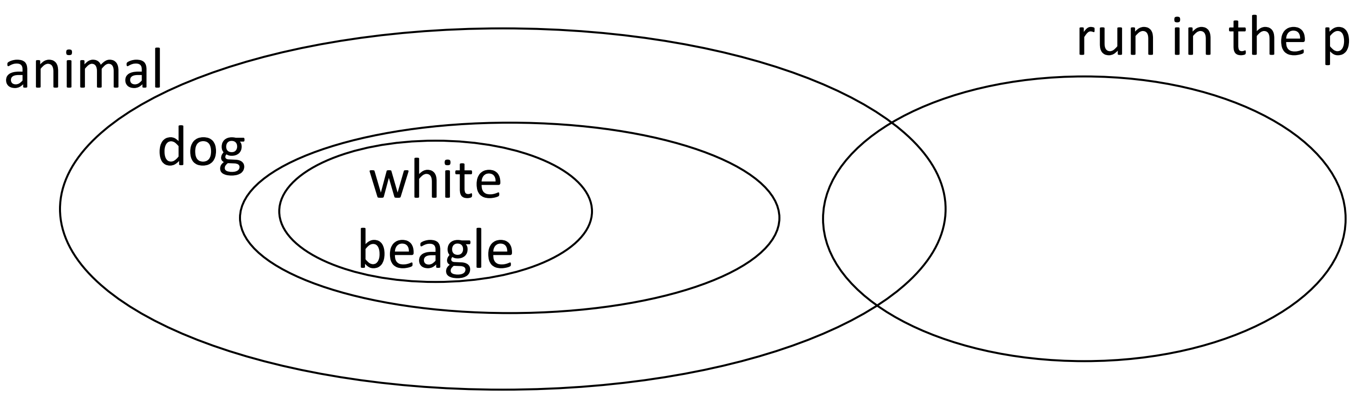
Monotonicity (Downward Monotone)

<u>Downward inferences</u>: **order reversing** inferences from general to specific phrases

P: **No** [dogs↓] ran in the park

H1: **No** [white beagles] ran in the park *<u>Entailment</u>*

H2: **No** [animals] ran in the park *<u>Non-Entailment</u>*



Monotonicity Reasoning

Upward inferences: inferences from specific to general phrases

P: **Some** [dogs↑] ran in the park

H: **Some** [animals] ran in the park *Entailment*

Downward inferences: inferences from general to specific phrases

P: **No** [dogs↓] ran in the park

H: **No** [white beagles] ran in the park *Entailment*

Question: Can current neural models compositionally capture various semantic phenomena for properly handling both directions of monotonicity reasoning?

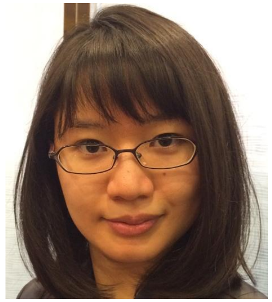
Previous NLI Datasets

- Previous datasets containing monotonicity inferences
  - FraCaS [Cooper+, 1994]: 37/346 examples
  - GLUE diagnostic dataset [Wang+, 2019]: 93/1,650 examples

  limited to very small sizes

- Standard NLI datasets for neural models
  - SNLI [Bowman+, 2015]
  - MultiNLI [Williams+, 2018]

  rarely come from monotonicity inference patterns



Hitomi Yanaka

Koji Mineshima

Daisuke Bekki

Satoshi Sekine

Kentaro Inui

Lasha Abzianidze

Johan Bos

MED

Monotonicity **E**ntailment **D**ataset for testing models on monotonicity reasoning [Yanaka+, BlackboxNLP2019]

https://github.com/verypluming/MED

MED: Monotonicity Entailment Dataset

[Yanaka+, BlackboxNLP2019] https://github.com/verypluming/MED

- Collect 5,382 examples including a wide range of monotonicity reasoning in two ways:

MED: Monotonicity Entailment Dataset

[Yanaka+, BlackboxNLP2019] https://github.com/verypluming/MED

- Collect 5,382 examples including a wide range of monotonicity reasoning in two ways:

- Human-oriented Dataset: 4,068 examples
  naturally-occurring inference examples by crowdsourcing

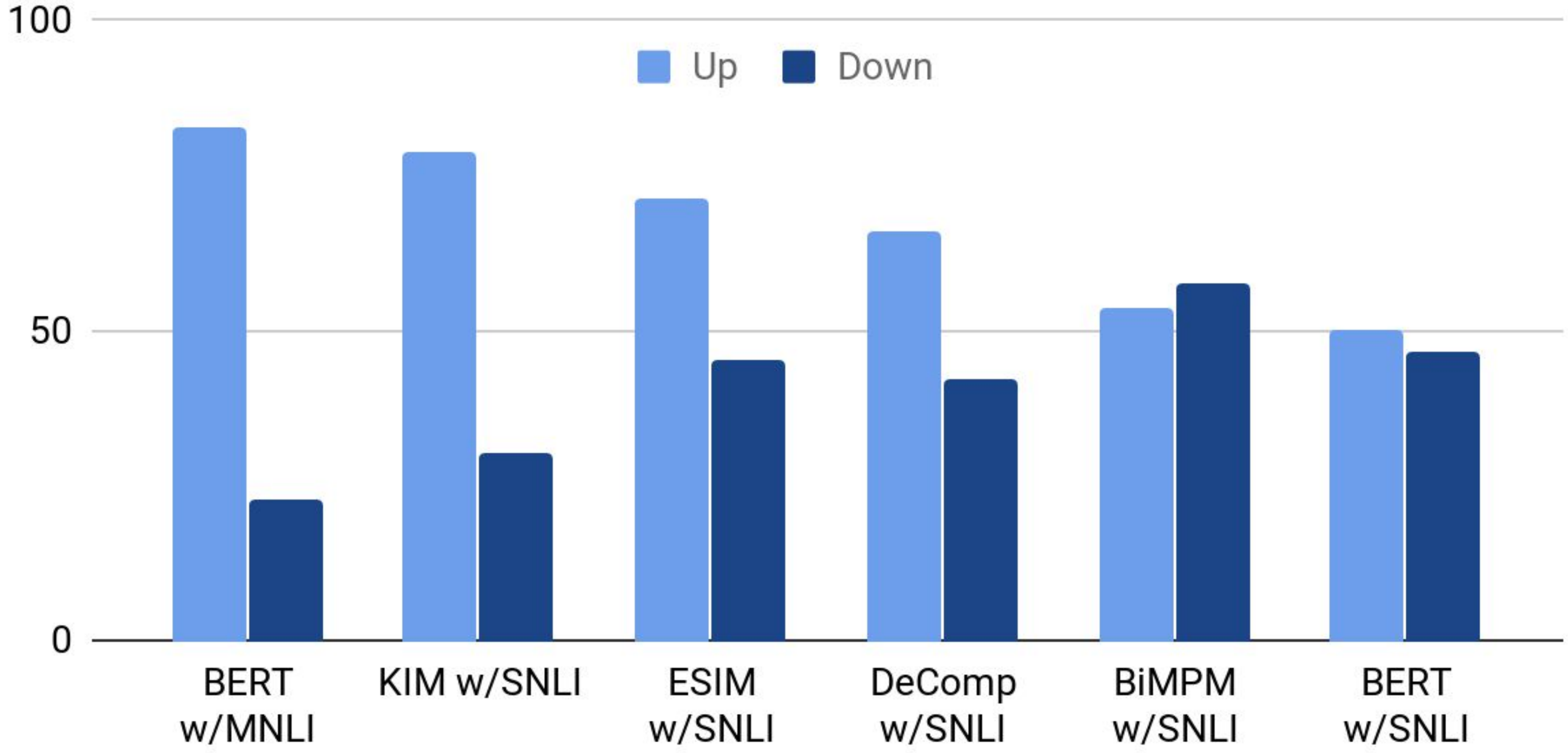
- Linguistics-oriented Dataset: 1,314 examples
  well-designed inference examples collected from 11 linguistics publications and previous NLI datasets (FraCaS and GLUE-diag)

Examples in MED

- upward (1,818)/downward (3,272)/non-monotone (292) (ccg2mono [Hu+,2018] + manual check)
- linguistic phenomena tags: lexical knowledge, conjunction, disjunction, conditionals, negative polarity items (NPI), reverse

Up	Lex	Human	P: He approached the boy reading **a** magazine H: He approached the boy reading **a** book *Entailment*
Up	Conj Rev	Human	P: I ca**n't** imagine a long life **without** music and cooking H: I ca**n't** imagine a long life **without** music *Entailment*
Down	Lex NPI	Human	P: Tom **hardly ever** listens to music H: Tom **hardly ever** listens to rock 'n' roll *Entailment*
Down	Disj	Paper	P: Almost **nobody** has had a sunburn or caught a cold H: Almost **nobody** has caught a cold *Entailment*

Performance of DNN models on MED



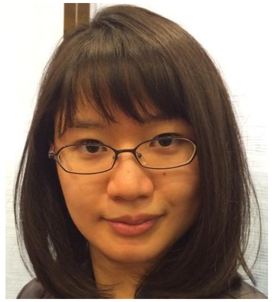
- DNN-based NLI models trained with benchmark datasets do not work well on downward monotonicity.
- The better a model performs on upward inferences, the worse it performs on downward inferences.

Possible reason for low performance on downward inferences: Lack of training datasets for downward inferences

Only 77/1700 examples in MultiNLI are downward inference examples involving the downward operator "no":

No racin' on the Range
No horse racing is allowed on the Range *Entailment*

Question:
Is the obstacle to downward inferences the size of training datasets?


Hitomi Yanaka


Koji Mineshima


Daisuke Bekki


Satoshi Sekine


Kentaro Inui


Lasha Abzianidze


Johan Bos

HELP

A Dataset for **H**andling **E**ntailments with
lexical and **L**ogical **P**henomena [Yanaka+, *SEM2019]

https://github.com/verypluming/HELP

HELP: A Dataset for Handling Entailments with lexical and Logical Phenomena [Yanaka+, *SEM2019]

https://github.com/verypluming/HELP

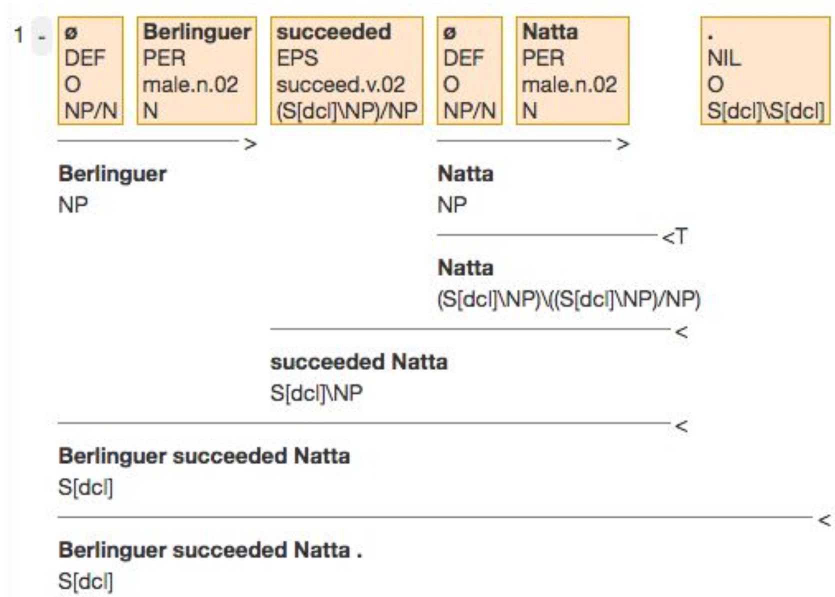
- HELP is an automatically generated monotonicity inference dataset that embodies the combination of lexical and logical inferences.
- We use HELP as a training set and MED for evaluation.
- We investigate whether automated data augmentation helps neural models to learn monotonicity reasoning.

Original Corpus: the Parallel Meaning Bank (PMB)

[Abzianidze+, 2017]

https://pmb.let.rug.nl/

- annotated with 72 types of semantic tags, word senses, Combinatory Categorial Grammar (CCG) [Steedman, 2000] syntactic analysis, and formal meaning representations
- manageable for our automatic creation of monotonicity inferences
  - monotone operators
  - syntactic structures
  - lexical knowledge



Automatic Monotonicity Dataset Creation

1. Select sentences including monotonicity properties (quantifiers, negation, conditionals, conjunction, disjunction) by using semantic tags

All	*kids*	*were*	*dancing*	*on*	*the*	*floor*
AND	CON	PST	EXG	REL	DEF	CON

Automatic Monotonicity Dataset Creation

1. Select sentences including monotonicity properties (quantifiers, negation, conditionals, conjunction, disjunction) by using semantic tags

All	*kids*	*were*	*dancing*	*on*	*the*	*floor*
AND	CON	PST	EXG	REL	DEF	CON

2. Determine the polarity of arguments by using CCG syntactic trees

P: **All** [$_{NP}$ **kids** ↓] were [$_{VP}$ dancing on the floor ↑]

Automatic Monotonicity Dataset Creation

1. Select sentences including monotonicity properties (quantifiers, negation, conditionals, conjunction, disjunction) by using semantic tags

All	*kids*	*were*	*dancing*	*on*	*the*	*floor*
AND	CON	PST	EXG	REL	DEF	CON

2. Determine the polarity of arguments by using CCG syntactic trees

P: **All** $[_{NP}$ **kids** $\downarrow]$ were $[_{VP}$ dancing on the floor $\uparrow]$

3. Replace words by using WordNet sense and create the hypothesis

P: **All** $[_{NP}$ **kids** $\downarrow]$ were $[_{VP}$ dancing on the floor $\uparrow]$

H: **All foster children** were dancing on the floor

Automatic Monotonicity Dataset Creation

1. Select sentences including monotonicity properties (quantifiers, negation, conditionals, conjunction, disjunction) by using semantic tags

All	*kids*	*were*	*dancing*	*on*	*the*	*floor*
AND	CON	PST	EXG	REL	DEF	CON

2. Determine the polarity of arguments by using CCG syntactic trees

P: **All** [$_{NP}$ **kids** ↓] were [$_{VP}$ dancing on the floor ↑]

3. Replace words by using WordNet sense and create the hypothesis

P: **All** [$_{NP}$ **kids** ↓] were [$_{VP}$ dancing on the floor ↑]

H: **All foster children** were dancing on the floor *Entailment*

Automatic Monotonicity Dataset Creation

1. Select sentences including monotonicity properties (quantifiers, negation, conditionals, conjunction, disjunction) by using semantic tags

All	*kids*	*were*	*dancing*	*on*	*the*	*floor*
AND	CON	PST	EXG	REL	DEF	CON

2. Determine the polarity of arguments by using CCG syntactic trees

P: **All** $[_{NP}$ **kids** ↓] were $[_{VP}$ dancing on the floor ↑]

3. Replace words by using WordNet sense and create the hypothesis

P: **All** $[_{NP}$ **kids** ↓] were $[_{VP}$ dancing on the floor ↑]

H: **All foster children** were dancing on the floor *Entailment*

4. Swap the premise and the hypothesis and create a new pair

P'(=H): **All** foster children were dancing on the floor

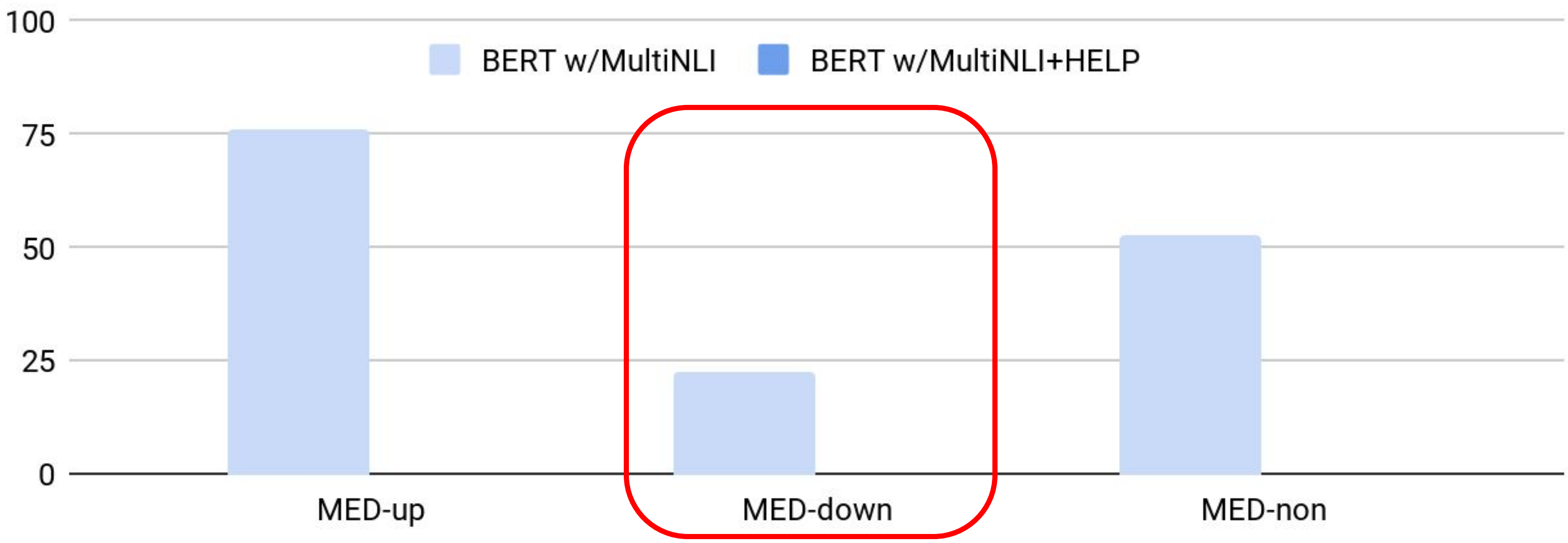
H'(=P): **All** kids were dancing on the floor *Non-entailment*

Examples in HELP

Total: 36K automatically generated inference pairs

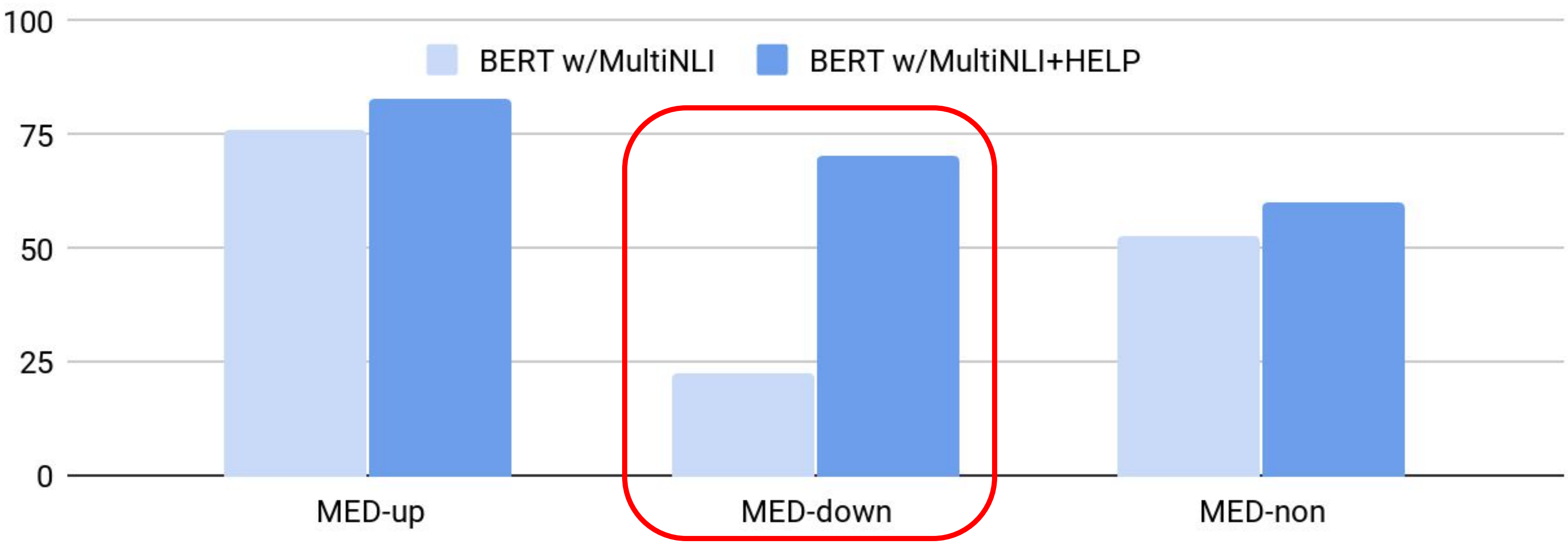
Section	Size	Example
upward monotone	7784	P: There are **some coneflowers** in the garden H: There are **some flowers** in the garden *Entail*
downward monotone	21192	P: In those days, there were **no radios** H: In those days, there were **no clock radios** *Entail*
non monotone	1105	P: Shakespeare wrote **both tragedy and comedy** H: Shakespeare wrote **both tragedy and drama** *Non-entail*
conjunction	6076	P: Tom removed his **glasses** H: Tom removed his **glasses and rubbed his eyes** *Non-entail*
disjunction	438	P: The trees **are barren** H: The trees **are barren or bear only small fruit** *Entail*

Performance of BERT on MED



BERT trained with only MultiNLI-train does not work well especially on downward monotonicity

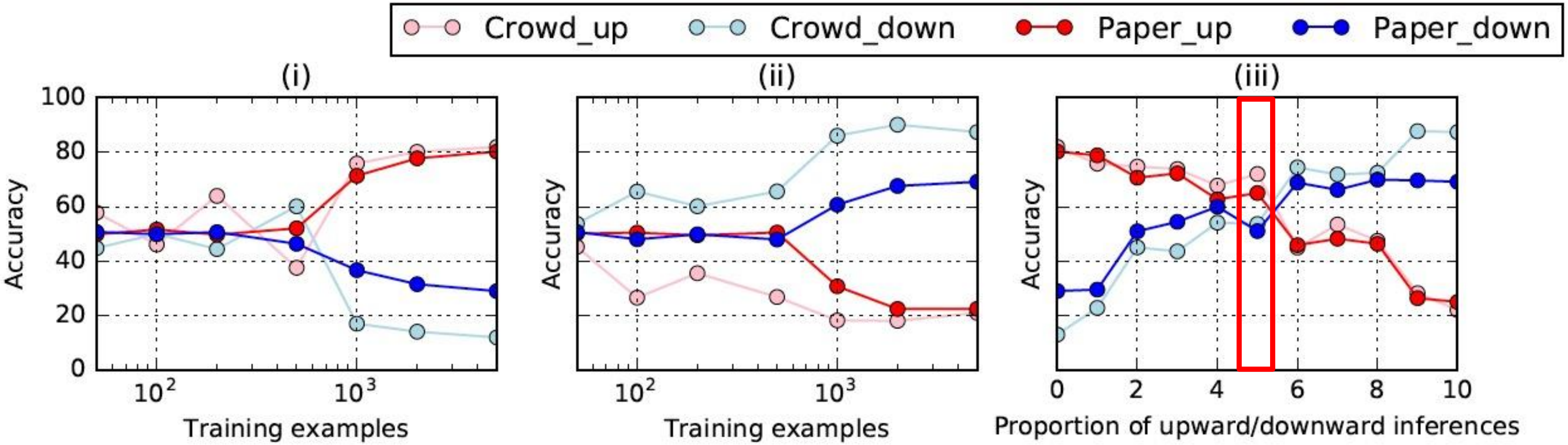
Performance of BERT on MED with HELP



HELP improved the performance of BERT on downward monotonicity

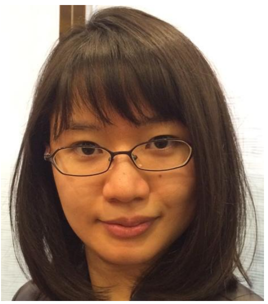
Relationship between Accuracy on Upward Inferences and Downward Inferences

Accuracy throughout training BERT with (i) only upward examples, (ii) only downward examples, and (iii) different ratios of upward/downward examples (Total: 5K examples)



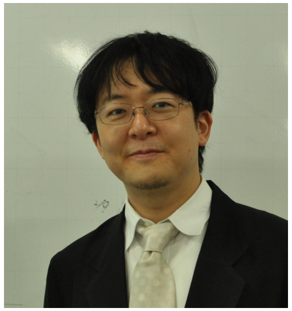
The performance depends on the majority distribution of a training set

Question: Do current models have limitations on their generalization ability in monotonicity reasoning?



Hitomi Yanaka

Koji Mineshima



Daisuke Bekki

Kentaro Inui

Systematicity

Do neural models learn systematicity of monotonicity inference in natural language? [Yanaka+, ACL2020]

https://github.com/verypluming/systematicity

Systematicity [Fodor and Pylyshin, 1988]

Systematicity: The ability to understand a sentence is connected to the ability to understand certain other sentences

Systematicity of Inference:

If you can infer from P&Q&R to P, you can also infer from P&Q to P

Systematicity [Fodor and Pylyshin, 1988]

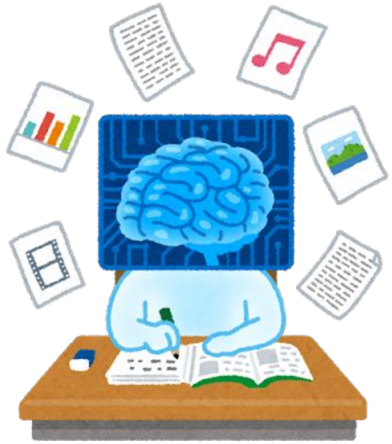
Systematicity: The ability to understand a sentence is connected to the ability to understand certain other sentences

Systematicity of Inference:

If you can infer from P&Q&R to P, you can also infer from P&Q to P

- If models obtain systematicity of inference, they should learn inferences from only a small number of training instances

Question:
Do neural models learn systematicity of inference in natural language?



Systematicity of Monotonicity

Upward inferences: inferences from specific to general phrases

P1: **Some** [small dogs↑] ran

H1: **Some** [dogs] ran

Entailment

Systematicity of Monotonicity

Upward inferences: inferences from specific to general phrases

P1: **Some** [small dogs↑] ran

H1: **Some** [dogs] ran

P2: **Several** [small dogs↑] ran

H2: **Several** [dogs] ran

Entailment

Systematicity of Monotonicity

Upward inferences: inferences from specific to general phrases

P1: **Some** [small dogs↑] ran

H1: **Some** [dogs] ran

P2: **Several** [small dogs↑] ran

H2: **Several** [dogs] ran

Entailment

Downward inferences: inferences from general to specific phrases

P3: **No** [dogs↓] ran

H3: **No** [beagles] ran

Entailment

Systematicity of Monotonicity

Upward inferences: inferences from specific to general phrases

P1: **Some** [small dogs↑] ran P2: **Several** [small dogs↑] ran

H1: **Some** [dogs] ran H2: **Several** [dogs] ran

Entailment

Downward inferences: inferences from general to specific phrases

P3: **No** [dogs↓] ran P4: **Few** [dogs↓] ran

H3: **No** [beagles] ran H4: **Few** [beagles] ran

Entailment

To handle monotonicity, models should systematically capture

1. **monotonicity direction of quantifiers (upward/downward)**

Systematicity of Monotonicity

Upward inferences: inferences from specific to general phrases

P1: **Some** [small dogs↑] ran

H1: **Some** [dogs] ran

P2: **Several** [small dogs↑] ran

H2: **Several** [dogs] ran

Entailment

Downward inferences: inferences from general to specific phrases

P3: **No** [dogs↓] ran

H3: **No** [beagles] ran

P4: **Few** [dogs↓] ran

H4: **Few** [beagles] ran

Entailment

To handle monotonicity, models should systematically capture

1. **monotonicity direction of quantifiers (upward/downward)**
2. **lexical and structural replacement (general/specific)**

Productivity of Monotonicity

If a propositional object is embedded in another downward context, the polarity of words over its scope can be reversed again

P: **All** [workers↓] joined for a French dinner
H: **All** [new workers] joined for a French dinner *Entailment*

P: **Not** [**all** [new workers↑]] joined for a French dinner
H: **Not** [**all** [workers]] joined for a French dinner *Entailment*

To handle monotonicity, models should systematically capture

1. **monotonicity direction of quantifiers (upward/downward)**
2. **lexical and structural replacement (general/specific)**
3. **productivity (recursiveness)**

Key idea of analyzing systematicity of neural models

To evaluate the systematic generalization ability of DNN-based NLI models on **monotonicity** and its **productivity**,
we propose a new evaluation protocol where we

1. **synthesize monotonicity inference datasets**
2. **systematically control which patterns are shown to the models during training and which are left unseen**

Synthesize Monotonicity Dataset

1. Generate a premise by using a context-free grammar

Examples of context-free grammar rules
N→{dogs, ...}, IV→{ran, ...}, TV→{chased, ...}, Q→{some, ...},
NP→Q N|Q N Sbar, S→ NP IV, Sbar→which TV NP

Synthesize Monotonicity Dataset

1. Generate a premise by using a context-free grammar

Examples of context-free grammar rules
N→{dogs, ...}, IV→{ran, ...}, TV→{chased, ...}, Q→{some, ...},
NP→Q N|Q N Sbar, S→ NP IV, Sbar→which TV NP

Some dogs ran (n=1)

Some dogs which chased **some** dogs ran (n=2)

Some dogs which chased **some** dogs which chased **some** dogs ran (n=3)

Synthesize Monotonicity Dataset

1. Generate a premise by using a context-free grammar

Examples of context-free grammar rules
N⟶{dogs, ...}, IV⟶{ran, ...}, TV⟶{chased, ...}, Q⟶{some, ...},
NP⟶Q N|Q N Sbar, S⟶ NP IV, Sbar⟶which TV NP

Some dogs ran (n=1)

Some dogs which chased **some** dogs ran (n=2)

Some dogs which chased **some** dogs which chased **some** dogs ran (n=3)

2. Rephrase the premise and generate hypotheses

P: **Some** [**dogs**] ran	H: **Some** [**animals**] ran	*Entailment*
	H': **Some** [**white dogs**] ran	*Non-entailment*

A context-free grammar and a set of phrase replacements

Context-free grammar for premise sentences		
S	$\rightarrow$	$NP\ IV_1$
NP	$\rightarrow$	$Q\ N \mid Q\ N\ \overline{S}$
$\overline{S}$	$\rightarrow$	$WhNP\ TV\ NP \mid WhNP\ NP\ TV \mid NP\ TV$
Lexicon		
Q	$\rightarrow$	{*no, at most three, less than three, few, some, at least three, more than three, a few*}
N	$\rightarrow$	{*dog, rabbit, lion, cat, bear, tiger, elephant, fox, monkey, wolf*}
IV_1	$\rightarrow$	{*ran, walked, came, waltzed, swam, rushed, danced, dawdled, escaped, left*}
IV_2	$\rightarrow$	{*laughed, groaned, roared, screamed, cried*}
TV	$\rightarrow$	{*kissed, kicked, hit, cleaned, touched, loved, accepted, hurt, licked, followed*}
$WhNP$	$\rightarrow$	{*that, which*}
N_{lex}	$\rightarrow$	{*animal, creature, mammal, beast*}
Adj	$\rightarrow$	{*small, large, crazy, polite, wild*}
PP	$\rightarrow$	{*in the area, on the ground, at the park, near the shore, around the island*}
$RelC$	$\rightarrow$	{*which ate dinner, that liked flowers, which hated the sun, that stayed up late*}
Adv	$\rightarrow$	{*slowly, quickly, seriously, suddenly, lazily*}
Phrase replacements for hypothesis sentences		
N	to	$N_{lex} \mid Adj\ N \mid N\ PP \mid N\ RelC$
IV_1	to	$IV_1\ Adv \mid IV_1\ PP \mid IV_1$ or $IV_2 \mid IV_1$ and IV_2

How to Test Systematicity

Train A Fix a quantifier
and feed various phrase replacements

Some puppies ran   **Some** white dogs ran

`Lex`   `Adj`

Some dogs ran

How to Test Systematicity

Train A Fix a quantifier and feed various phrase replacements

Train B Fix a phrase replacement and feed various quantifiers

Some puppies ran $\xrightarrow{\texttt{Lex}}$ **Some** dogs ran $\xleftarrow{\texttt{Adj}}$ **Some** white dogs ran

Several puppies ran $\xrightarrow{\texttt{Lex}}$ **Several** dogs ran

No dog ran $\xrightarrow{\texttt{Lex}}$ **No** puppie ran

How to Test Systematicity

<u>Train A</u> Fix a quantifier and feed various phrase replacements

Some <u>puppies</u> ran **Some** <u>white dogs</u> ran

Lex Adj

Some <u>dogs</u> ran

<u>Train B</u> Fix a phrase replacement and feed various quantifiers

Several <u>puppies</u> ran **No** <u>dog</u> ran

Lex Lex

Several <u>dogs</u> ran **No** <u>puppie</u> ran

<u>Test</u> Unseen combinations of quantifiers and phrase replacements

Several <u>white dogs</u> ran

Adj

Several <u>dogs</u> ran

No <u>white dogs</u> ran

Adj

No <u>dogs</u> ran

How to Test Productivity

Train A Depth 1

Some puppies ran

↓ {Lex,Adj,Prep,…}

Some dogs ran

Train B Depth 2

Some dogs which chased **some** puppies ran

↓ {Lex,Adj,Prep,…}

Some dogs which chased **some** dogs ran

How to Test Productivity

Train A Depth 1

Some puppies ran

↓ **{Lex,Adj,Prep,…}**

Some dogs ran

Train B Depth 2

Some dogs which chased **some** puppies ran

↓ **{Lex,Adj,Prep,…}**

Some dogs which chased **some** dogs ran

Test Unseen depths

Some dogs which chased **some** dogs which followed **some** puppies ran

↓ **{Lex,Adj,Prep,…}**

Some dogs which chased **some** dogs which followed **some** dogs ran

Experimental Setting

- Models
  - LSTM [Hochreiter and Schmidhuber, 1997]
  - TreeLSTM [Tran and Cheng, 2018]
  - BERT-based NLI [Devlin+, 2018]

- Datasets
  - Train/Test = 300,000/20,000
  - *Entailment*:*Non-entailment* = 1:1 (Chance rate: 0.5)
  - Upward:Downward = 1:1

- Evaluation metrics: the average accuracy of 5 runs

Experiment 1: Systematicity

Train

Train A. 1 quantifier × All replacements

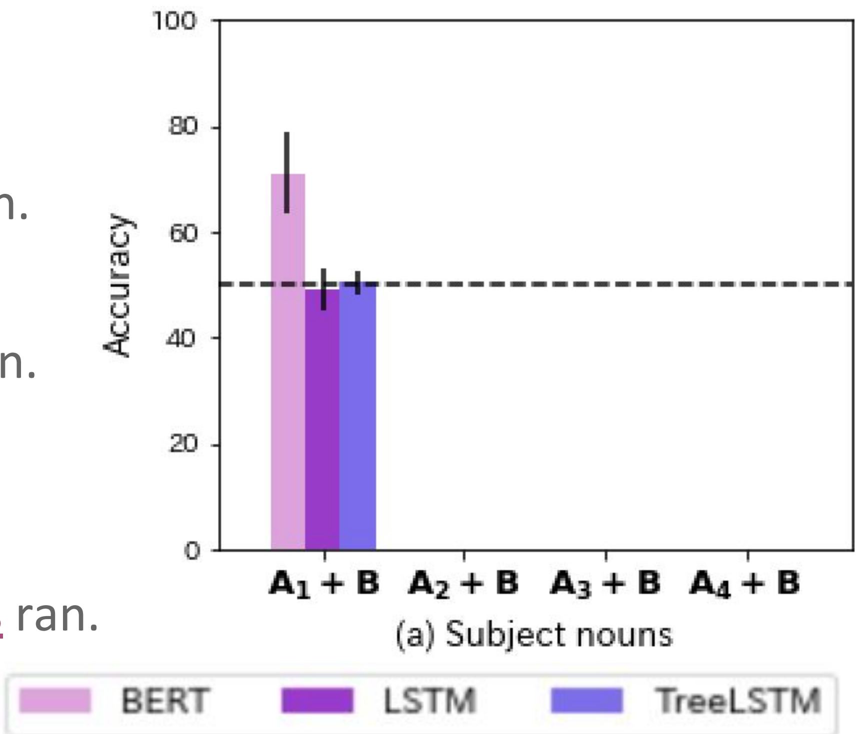
P: **Some** puppies ran. ⟶ H: **Some** dogs ran.

P: **Some** white dogs ran. ⟶ H: **Some** dogs ran.

Train B. All quantifiers × 1 replacement

P: **Several** puppies ran. ⟶ H: **Several** dogs ran.

P: **No** dogs ran. ⟶ H: **No** puppies ran.



Test Unseen combinations

P: **Several** white dogs ran. ⟶ H: **Several** dogs ran.

P: **No** dogs ran. ⟶ H: **No** white dogs ran.

- BERT generalizes to unseen combinations of quantifiers and phrase replacements

Experiment 1: Systematicity

Train Gradually add **Train A** to the training set

Train A. 1 quantifier × All replacements

P: **Some** puppies ran. ⟶ H: **Some** dogs ran.

P: **Some** white dogs ran. ⟶ H: **Some** dogs ran.

Train B. All quantifiers × 1 replacement

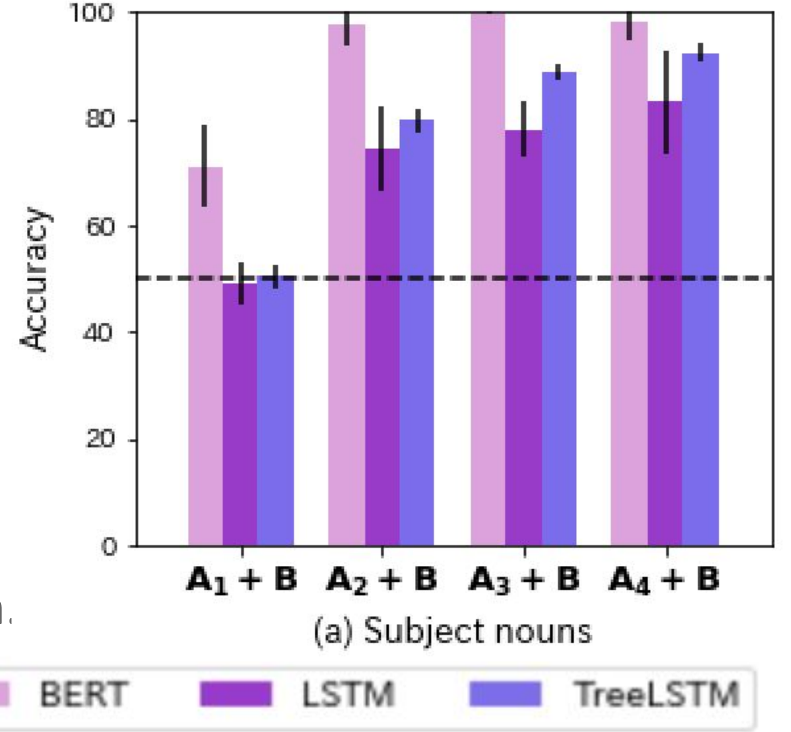
P: **Several** puppies ran. ⟶ H: **Several** dogs ran.

P: **No** dogs ran. ⟶ H: **No** puppies ran.

Test Unseen combinations

P: **Several** white dogs ran. ⟶ H: **Several** dogs ran.

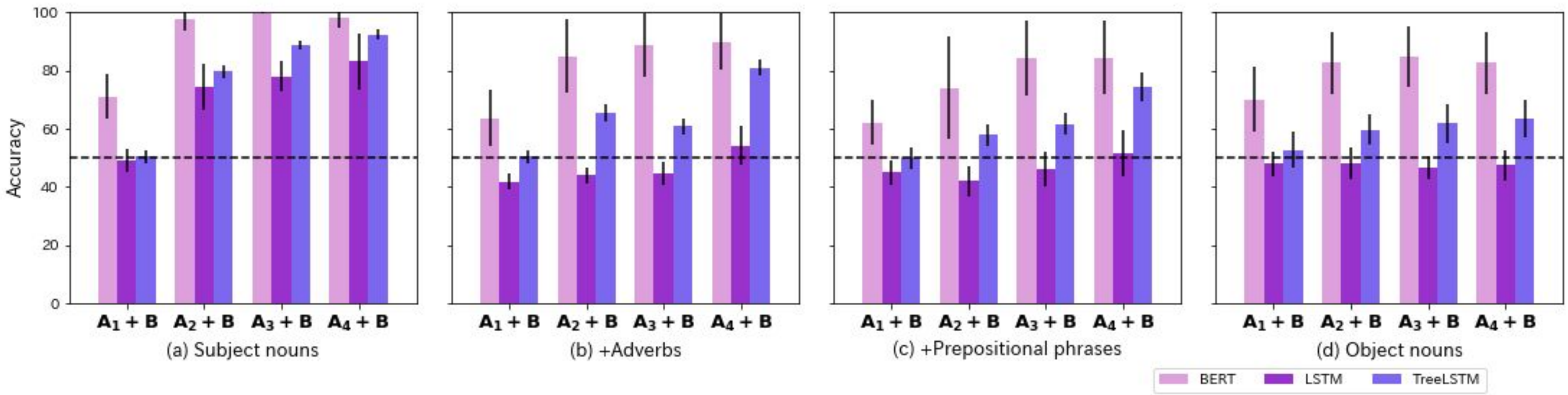
P: **No** dogs ran. ⟶ H: **No** white dogs ran.



- BERT generalizes to unseen combinations of quantifiers and phrase replacements
- The accuracy is better as more training data are fed into models

Experiment 1: Systematicity

When testing models on slightly different syntactic structures:



P: **A** dog ran
H: **An** animal ran
Entail

P’: Today **a** dog ran
H’: Today **an** animal ran
Entail

P’’: In the park, **a** dog ran
H’’: In the park, **an** animal ran
Entail

P’’’: I saw **a** dog
H’’’: I saw **an** animal
Entail

- The accuracy of all models significantly decreased
- This decrease becomes larger as the syntactic structures in the test set become different from those in the training set

Experiment 2: Productivity

Train	Dev/Test	BERT	LSTM	TreeLSTM
depth1 + depth2	depth1	100	100	100
	depth2	100	99.8	99.5
	depth3	75.2	75.4	86.4
	depth4	55.9	57.7	58.6
	depth5	49.9	45.8	48.4
depth1 + depth2 + depth3	depth1	100	100	100
	depth2	100	95.1	99.6
	depth3	100	85.2	97.7
	depth4	77.9	59.7	68.0
	depth5	53.5	55.1	49.6

- All models generalize to one level deeper depth
- But they fail to generalize to two level deeper

When MultiNLI [Williams+2018] is Added to the Training Set

Train	Dev/Test	BERT	LSTM	TreeLSTM
MNLI	depth1	46.9	47.2	43.4
	depth2	46.2	48.3	49.5
	depth3	46.8	48.9	41.0
	depth4	48.5	50.6	48.5
	depth5	48.9	49.3	48.8
	MNLI	84.6	64.7	70.4
depth1 + depth2 + MNLI	depth1	100	100	100
	depth2	100	89.3	99.8
	depth3	67.8	66.7	76.3
	depth4	46.8	47.1	50.7
	depth5	41.2	46.7	47.5
	MNLI	84.4	39.7	63.0

- Only the BERT maintains the performance on MultiNLI while improving the performance on monotonicity inferences

When MultiNLI [Williams+2018] is Added to the Training Set

Train	Dev/Test	BERT	LSTM	TreeLSTM
MNLI	depth1	46.9	47.2	43.4
	depth2	46.2	48.3	49.5
	depth3	46.8	48.9	41.0
	depth4	48.5	50.6	48.5
	depth5	48.9	49.3	48.8
	MNLI	84.6	64.7	70.4
depth1 + depth2 + MNLI	depth1	100	100	100
	depth2	100	89.3	99.8
	depth3	67.8	66.7	76.3
	depth4	46.8	47.1	50.7
	depth5	41.2	46.7	47.5
	MNLI	84.4	39.7	63.0

- But all models still fail to generalize to two level deeper

Conclusion

Motivation

Evaluating whether DNN models can learn the compositional generalization capacity underlying NLI

Main results

- The generalization ability of DNN models is limited to cases where the syntactic structures are similar to those in the training set
- BERT might have the ability to memorize different types of datasets

Future Work

- Investigating how to improve the generalization capacity of DNN models
  - Data augmentation, Multi-task learning, Architecture refinement

Thanks!

Hitomi Yanaka hitomi.yanaka@riken.jp
http://hitomiyanaka.strikingly.com/