JaNLI: 日本語の言語現象に基づく敵対的推論データセット

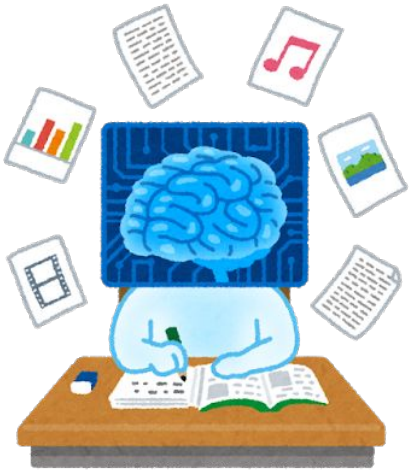
谷中　瞳[1]、峯島　宏次[2]
[1]東大、[2]慶応大

汎用言語モデルによる言語理解

- 深層ニューラルネット(Deep Neural Network)による事前学習に基づく汎用言語モデルが活発に研究されている
  - BERT[Devlin+ 18], T5[Raffel+ 19], GPT-3[Brown+ 20]

- 高度な言語理解タスクの大規模ベンチマークにおいて高性能を達成しつつある
  - GLUE[Wang+ 18], SuperGLUE[Wang+ 19]

汎用言語モデルによる言語理解の可能性？



自然言語推論 (Natural Language Inference, NLI)

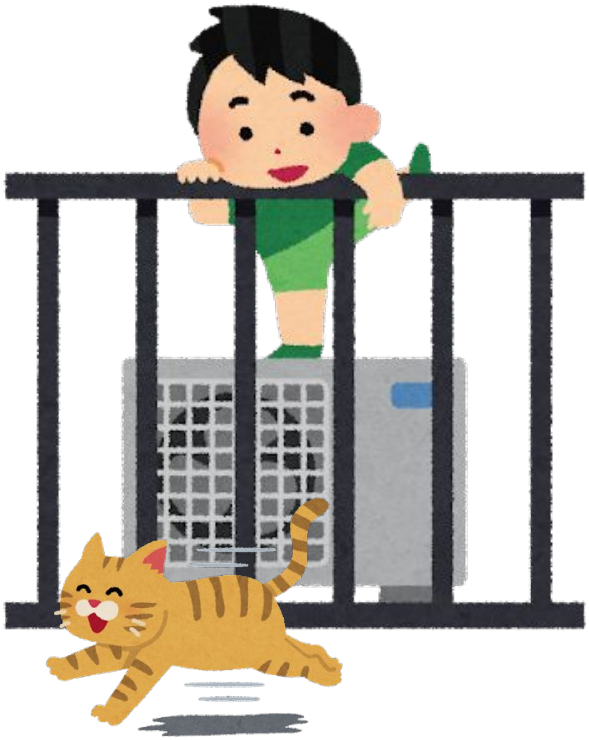
含意関係認識 (Recognizing Textual Entailment, RTE)とも
コンピュータによるテキスト間の言語理解に向けたタスク
前提文に対して仮説文は同じ意味を含むか（含意関係）

前提文　子供が走っている猫を見ている
仮説文　猫が走っている

含意

前提文　子供が走っている猫を見ている
仮説文　子供が走っている

非含意



NLIデータセットの関連研究

- 英語は多種多様なNLIデータセットが存在
  - 言語学者による構築：FraCaS[Cooper 94]
  - クラウドソーシングによる構築：
    SNLI[Bowman+ 15]、MultiNLI[Williams+ 18]
    SICK[Marelli+ 14], SemEval2012-2017
- 近年、多言語化が進む
  - MultiNLI：XNLI(15ヶ国語)[Conneau+ 18], 韓国語[Ham+ 20]
  - SICK：ポルトガル語[Real+ 18], オランダ語[Wijnholds+ 18]
- その中で日本語は発展途上
  - JSeM[Kawazoe+ 17]: 言語学者によるFraCaSの日本語版
  - JSICK[谷中&峯島 21]: SICKを人手で翻訳＋クラウド
  - JSNLI[吉越+ 20]: SNLIを機械翻訳+自動フィルタ+クラウド
  - 旅行口コミを用いた根拠付RTEデータセット[Hayashibe 20]

HANS (Heuristic Analysis for NLI Systems)

[McCoy+ 2019]

深層学習のモデルが人のように単語の意味と文構造に従って、様々な文の意味を**構成的に**理解しているか評価する目的で構築された、英語のNLIデータセット

- モデルが陥りやすい3つのヒューリスティクスを定義
  - ヒューリスティクスに従うと非含意のケースを含意と誤判定
- 言語現象に基づくテンプレートを設計し、自動構築

Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	**The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor.
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept.

目的: 日本語の言語現象を用いた深層学習モデルのヒューリスティクス分析

- 英語のHANSデータセットを参考に、モデルが陥りやすいヒューリスティクスごとに日本語の言語現象に基づく推論テンプレートを設計し、推論データセットを自動構築

- 日本語・多言語の汎用言語モデルがどのくらい日本語の統語・意味的知識に基づいて構成的に推論できているのか分析

日本語の言語現象に基づく敵対的推論データセット JaNLI[Yanaka&Mineshima,BlackboxNLP2021]の特徴

ヒューリスティクスごとに日本語の言語現象に基づく推論テンプレートを設計し、推論データセットを自動構築

1. 日本語の柔軟な語順を考慮して5つのヒューリスティクスを定義

2. ガーデンパス現象を含め16の日本語の言語現象を用いて144の含意・非含意の推論テンプレートを設計

1．推論の5つのヒューリスティクス

英語HANSの3つのヒューリスティクス (subsequence, constituent, lexical overlap)を参考に、日本語の推論の5つのヒューリスティクスを定義

subsequence	男の子 が 眠っている 女の子 を 見ている 男の子 が 眠っている 非含意
constituent	ひょっとしたら 子供 が 眠っている 子供 が 眠っている 非含意
full-overlap	ライダー が サーファー を 助け出した ライダー を サーファー が 助け出した 非含意
order-preserving subset	学生 か 子供 が 遊んでいる 学生 が 遊んでいる 非含意
mixed-subset	子供 が 泳いでいる 学生 を 助け出した 子供 を 学生 が 助け出した 非含意

1. 推論の5つのヒューリスティクス

日本語JaNLIでは語順の違いを考慮して、lexical overlapを次の3通りに細分化

full-overlap：前提文と仮説文で全単語共通・語順異なる
order-preserving subset：仮説文の語を含む・語順同じ
mixed-subset：仮説文の語を含む・語順異なる

full-overlap	ライダー が サーファー を 助け出した ライダー を サーファー が 助け出した　非含意
order-preserving subset	学生 か 子供 が 遊んでいる 学生 が 遊んでいる　非含意
mixed-subset	子供 が 泳いでいる 学生 を 助け出した 子供 を 学生 が 助け出した　非含意

2. 日本語の16の言語現象

スクランブリング、受身、使役、事実性をはじめ、日本語の16の言語現象に基づいて、144の推論テンプレートを設計

Templates for *P* and *H*	Example	Phenomenon/Pattern
	子供 が 女性 を 見ている	
P: NP1 ga NP2 o TV-o	child ga woman o looking	Scrambling
	(*The child is looking at the woman*)	
⇒ H_1: NP2 o NP1 ga TV-o	女性 を 子供 が 見ている (*The child is looking at the woman*)	FULL-OVERLAP
⇏ H_2: NP1 o NP2 ga TV-o	子供 を 女性 が 見ている (*The woman is looking at the child*)	FULL-OVERLAP
⇏ H_3: NP2 ga NP1 o TV-o	女性 が 子供 を 見ている (*The woman is looking at the child*)	FULL-OVERLAP
	男の子 が 若者 に 押された	
P: NP1 ga NP2 ni TV-o passive	boy ga young-man ni push-passive	Passive
	(*The boy was pushed by the young man*)	
⇏ H_1: NP1 ga NP2 o TV-o	男の子 が 若者 を 押した (*The boy pushed the young man*)	ORDER-SUBSET
⇒ H_2: NP2 ga NP1 o TV-o	若者 が 男の子 を 押した (*The young man pushed the boy*)	MIXED-SUBSET
	男の子 が カップル を 笑わせている	
P: NP1 ga NP2 o IV causative	boy ga couple o laugh-causative	Causative
	(*The boy is making the couple laugh*)	
⇏ H_1: NP1 ga IV	男の子 が 笑っている (*The boy is laughing*)	ORDER-SUBSET
⇒ H_2: NP2 ga IV	カップル が 笑っている (*The couple is laughing*)	ORDER-SUBSET
	もしかしたら サーファー が 泳いでいる	
P: Factive-adverb NP1 ga IV	perhaps surfer ga swimming	Factive adverb
	(*Perhaps the surfer is swimming*)	
⇏ H_1: NP1 ga IV	サーファー が 泳いでいる	CONSTITUENT
	(*The surfer is swimming*)	

ガーデンパス現象

- ガーデンパス文
  文の解釈の途中で再解釈が必要となる文
  計算心理言語学を中心に読み時間の分析に用いられる
- 推論のパフォーマンスにも影響があるのか、人とモデルの性能を比較

Templates for P and H	Sentence Example	Phenomenon/Pattern
	子供 が 走っている 猫 を 追いかけた	
P: NP1 ga IV NP2 o TV-o	child ga running cat o chased	Garden-path sentence
	(*The child chased the running cat*)	
$\not\Rightarrow H_1$: NP1 ga IV	子供 が 走っている (*The child is running*)	SUBSEQUENCE
$\Rightarrow H_2$: NP2 ga IV	猫 が 走っている (*The cat is running*)	MIXED-SUBSET
$\Rightarrow H_3$: NP1 ga NP2 o TV-o	子供 が 猫 を 追いかけた (*The child chased the cat*)	ORDER-SUBSET
$\not\Rightarrow H_4$: NP1 o NP2 ga TV-o	子供 を 猫 が 追いかけた (*The cat chased the child*)	MIXED-SUBSET

ガーデンパス現象

日本語のガーデンパス文は人でも間違えて解釈しやすいが、
読点の有無など解釈を簡単にするファクターがいくつかある
解釈を簡単にするファクター別に5つのサブカテゴリを用意

Subcategory	Template	Example
GP-double-o	NP1 ga NP2 o TV-o1 NP3 o TV-o2	子供 が 猫 を 助けた 女の子 を 追いかけた child ga cat o rescued girl o chased (*The child chased the girl who rescued the cat*)
GP-punctuation	NP1 ga , IV NP2 o TV-o	子供 が , 走っている 猫 を 追いかけた child ga PUNCT running cat o chased (*The child chased the running cat*)
GP-selectional	NP-non-human ga IV-human NP2 o TV-o	リス が しゃべっている 女性 を 追いかけた squirrel ga talking woman o chased (*The squirrel chased the woman who was talking*)
GP-wa	NP1 wa IV NP2 o TV-o	子供 は 走っている 猫 を 追いかけた child wa running cat o chased (*The child chased the running cat*)
GP-scrambling	IV NP2 o NP1 ga TV-o	走っている 猫 を 子供 が 追いかけた running cat o child ga chased (*The child chased the running cat*)

推論テンプレートを用いたJaNLIの自動構築

144の推論テンプレートに対して、JSICK, JSNLIで20回以上出現する158語をランダムに割り当て、100件ずつ自動構築

Pattern (Heuristics)	Entailment	Non-entailment	Total
FULL-OVERLAP	800	1,200	2,000
ORDER-SUBSET	1,600	800	2,400
MIXED-SUBSET	3,400	2,000	5,400
SUBSEQUENCE	200	2,000	2,200
CONSTITUENT	1,200	1,200	2,400
Total	7,200	7,200	14,400

Linguistic Phenomenon	Examples (Templates)
GP-normal	1,600 (16)
GP-double-o	800 (8)
GP-punctuation	800 (8)
GP-selectional	800 (8)
GP-wa	800 (8)
GP-scrambling	1,600 (16)
Scrambling	1,600 (16)
Passive	400 (4)
Causative	400 (4)
Factive adverb	800 (8)
Factive verb	800 (8)
Modal	600 (6)
Negation	600 (6)
NP-coordination	1,200 (12)
Sentence-subordination	800 (8)
Sentence-coordination	800 (8)
Total	14,400 (144)

ベースライン実験

- NLIを含め様々な言語理解タスクで高精度の事前学習済み言語モデルBERT[Devlin+ 19]の日本語・多言語版をベースラインとして評価
- JaNLI720件について、クラウドソーシングで人の推論のパフォーマンスも評価し、モデルと比較
- BERTの実験設定
  - huggingfaceのモデルを使用
  - 4条件の学習データで、含意・非含意の2値分類タスクとしてファインチューニングし、正答率を評価
    (a)JSICK, (b)JSNLI
    (c)JSICK+JaNLI（一部）, (d)JSNLI+JaNLI（一部）
  - JSICKとJSNLIは含意・矛盾・中立の3値ラベルのため、矛盾・中立を非含意として扱った

評価結果（ヒューリスティクス別）

Model	Finetuned on	Correct: *Entailment*					Correct: *Non-entailment*				
		Full.	Order.	Mixed.	Subseq.	Const.	Full.	Order.	Mixed.	Subseq.	Const.
Ja	JSICK (5K)	99.9±0.00	97.8±0.02	79.4±0.10	98.3±0.02	88.6±0.07	0.1±0.00	6.2±0.01	6.7±0.04	32.5±0.11	22.7±0.09
	+JaNLI (0.7K)	90.8±0.04	98.6±0.01	96.8±0.02	99.2±0.01	97.3±0.02	67.1±0.17	59.1±0.04	84.6±0.23	92.4±0.09	90.4±0.05
	JSNLI (533K)	98.6±0.02	99.0±0.01	97.2±0.02	97.7±0.02	99.6±0.00	6.8±0.06	4.6±0.04	2.6±0.03	1.1±0.02	0.1±0.00
	+JaNLI (0.7K)	71.7±0.03	88.4±0.03	81.4±0.07	85.0±0.16	92.5±0.05	53.4±0.07	46.6±0.10	69.2±0.16	48.5±0.03	67.9±0.25
Multi	JSICK (5K)	66.0±0.57	64.6±0.56	57.1±0.50	62.7±0.55	63.8±0.55	33.9±0.57	34.7±0.57	36.2±0.55	45.1±0.48	43.5±0.49
	+JaNLI (0.7K)	40.8±0.37	32.9±0.33	38.0±0.35	49.8±0.44	38.8±0.36	64.2±0.33	66.0±0.37	83.3±0.19	77.4±0.32	80.9±0.23
	JSNLI (533K)	99.0±0.01	99.2±0.01	97.3±0.01	98.8±0.01	99.2±0.01	2.0±0.02	1.6±0.01	0.8±0.01	1.2±0.01	0.8±0.01
	+JaNLI (0.7K)	26.4±0.46	30.4±0.53	28.0±0.49	26.7±0.46	28.4±0.49	79.4±0.36	76.9±0.40	82.4±0.30	26.7±0.46	79.0±0.36
Human		94.2±0.05	97.1±0.01	92.7 ±0.04	100.0±0.00	98.3±0.03	97.8±0.01	95.8±0.05	88.7±0.09	94.3±0.08	91.1±0.14

- 人はほぼ完璧にできている非含意関係の推論を、日本語・多言語BERTは正しく推論できていない
  - 人もモデルもmixed subsetが低い傾向
- JaNLIを一部学習に追加したとき：
  - JaNLIだけでなく、JSICK, JSNLIの正答率も向上する傾向
  - 日本語BERTより多言語BERTの方が正答率が向上しにくい傾向

評価結果（ヒューリスティクス別）

Model	Finetuned on	Correct: *Entailment*					Correct: *Non-entailment*				
		Full.	Order.	Mixed.	Subseq.	Const.	Full.	Order.	Mixed.	Subseq.	Const.
Ja	JSICK (5K)	99.9±0.00	97.8±0.02	79.4±0.10	98.3±0.02	88.6±0.07	0.1±0.00	6.2±0.01	6.7±0.04	32.5±0.11	22.7±0.09
	+JaNLI (0.7K)	90.8±0.04	98.6±0.01	96.8±0.02	99.2±0.01	97.3±0.02	67.1±0.17	59.1±0.04	84.6±0.23	92.4±0.09	90.4±0.05
	JSNLI (533K)	98.6±0.02	99.0±0.01	97.2±0.02	97.7±0.02	99.6±0.00	6.8±0.06	4.6±0.04	2.6±0.03	1.1±0.02	0.1±0.00
	+JaNLI (0.7K)	71.7±0.03	88.4±0.03	81.4±0.07	85.0±0.16	92.5±0.05	53.4±0.07	46.6±0.10	69.2±0.16	48.5±0.03	67.9±0.25
Multi	JSICK (5K)	66.0±0.57	64.6±0.56	57.1±0.50	62.7±0.55	63.8±0.55	33.9±0.57	34.7±0.57	36.2±0.55	45.1±0.48	43.5±0.49
	+JaNLI (0.7K)	40.8±0.37	32.9±0.33	38.0±0.35	49.8±0.44	38.8±0.36	64.2±0.33	66.0±0.37	83.3±0.19	77.4±0.32	80.9±0.23
	JSNLI (533K)	99.0±0.01	99.2±0.01	97.3±0.01	98.8±0.01	99.2±0.01	2.0±0.02	1.6±0.01	0.8±0.01	1.2±0.01	0.8±0.01
	+JaNLI (0.7K)	26.4±0.46	30.4±0.53	28.0±0.49	26.7±0.46	28.4±0.49	79.4±0.36	76.9±0.40	82.4±0.30	26.7±0.46	79.0±0.36
Human		94.2±0.05	97.1±0.01	92.7 ±0.04	100.0±0.00	98.3±0.03	97.8±0.01	95.8±0.05	88.7±0.09	94.3±0.08	91.1±0.14

- 人はほぼ完璧にできている非含意関係の推論を、日本語・多言語BERTは正しく推論できていない
  - 人もモデルもmixed subsetが低い傾向
- JaNLIを一部学習に追加したとき：
  - JaNLIだけでなく、JSICK, JSNLIの正答率も向上する傾向
  - 日本語BERTより多言語BERTの方が正答率が向上しにくい傾向

評価結果（ヒューリスティクス別）

Model	Finetuned on	Test-overall	
		In-dist.	JaNLI
Ja	JSICK (5K)	92.1±0.01	51.3±0.01
	+JaNLI (0.7K)	92.3±0.01	89.3±0.06
	JSNLI (533K)	94.5±0.00	50.4±0.00
	+JaNLI (0.7K)	95.5±0.00	72.3±0.01
Multi	JSICK (5K)	73.6±0.20	50.2±0.01
	+JaNLI (0.7K)	86.5±0.08	56.9±0.06
	JSNLI (533K)	94.6±0.01	49.7±0.00
	+JaNLI (0.7K)	94.8±0.01	56.3±0.09
Human		-	94.0±0.04

- 人はほぼ完璧にできている非含意関係の推論を、日本語・多言語BERTは正しく推論できていない
  - 人もモデルもmixed subsetが低い傾向
- JaNLIを一部学習に追加したとき：
  - JaNLIだけでなく、JSICK, JSNLIの正答率も向上する傾向
  - 日本語BERTより多言語BERTの方が正答率が向上しにくい傾向

評価結果（言語現象別）

Model	Finetuned on	GP	Scramb.	Pass.	Caus.	Fac-adv.	Fac-v.	Modal	Neg.	NP-coord.	Subord.	Sent-coord.
Ja	JSICK	49.3±0.01	50.1±0.00	49.6±0.01	47.7±0.03	49.7±0.00	51.1±0.02	54.8±0.04	63.2±0.03	50.2±0.00	69.3±0.02	46.8±0.02
	+JaNLI	92.8±0.10	79.2±0.06	49.2±0.01	56.1±0.00	75.7±0.10	90.0±0.07	93.7±0.07	98.6±0.02	99.0±0.01	98.4±0.01	97.8±0.01
	JSNLI	50.2±0.01	52.3±0.02	45.9±0.04	49.7±0.01	51.5±0.01	51.2±0.01	49.6±0.00	50.2±0.01	51.4±0.00	50.0±0.00	49.7±0.00
	+JaNLI	70.1±0.06	65.3±0.03	41.2±0.06	50.5±0.01	67.9±0.08	70.2±0.09	71.7±0.19	87.4±0.06	76.6±0.17	88.8±0.11	79.2±0.18
Multi	JSICK	49.3±0.01	49.9±0.00	49.6±0.01	48.6±0.02	49.5±0.01	50.8±0.01	50.5±0.01	49.3±0.01	49.8±0.00	61.0±0.10	49.6±0.01
	+JaNLI	56.3±0.05	52.7±0.03	49.2±0.01	56.0±0.06	53.2±0.04	58.7±0.09	57.6±0.20	62.7±0.24	61.0±0.12	61.5±0.10	60.7±0.10
	JSNLI	49.8±0.00	50.1±0.00	48.1±0.01	49.9±0.00	50.3±0.00	50.3±0.00	49.6±0.01	45.5±0.04	50.5±0.00	49.9±0.00	50.2±0.00
	+JaNLI	54.1±0.07	53.8±0.07	48.9±0.02	50.7±0.01	52.7±0.05	53.3±0.06	55.3±0.09	62.6±0.22	54.4±0.08	54.4±0.08	54.8±0.08
Human		94.2±0.05	93.3±0.03	91.7±0.08	85.0±0.17	95.8±0.05	95.0±0.02	95.6±0.08	94.4±0.05	93.9±0.03	96.7±0.04	92.5±0.09

- JaNLIを一部学習に追加したとき
  - 多言語BERTの方が正答率が向上しにくい傾向
  - 日本語BERTも、スクランブリング、受身、使役、事実性副詞の正答率は向上しにくい傾向

JaNLIを学習に追加しても解けなかった推論の例

- スクランブリング、受身、使役、事実性副詞のケース
  - 語順や助詞、語の繰り返しはデータ拡張では捉えるのが困難？

Templates for P and H	Example	Phenomenon/Pattern
	子供 が 女性 を 見ている	
P: NP1 ga NP2 o TV-o	child ga woman o looking	Scrambling
	(*The child is looking at the woman*)	
$\Rightarrow H_1$: NP2 o NP1 ga TV-o	女性 を 子供 が 見ている (*The child is looking at the woman*)	FULL-OVERLAP
$\not\Rightarrow H_2$: NP1 o NP2 ga TV-o	子供 を 女性 が 見ている (*The woman is looking at the child*)	FULL-OVERLAP
$\not\Rightarrow H_3$: NP2 ga NP1 o TV-o	女性 が 子供 を 見ている (*The woman is looking at the child*)	FULL-OVERLAP
	男の子 が 若者 に 押された	
P: NP1 ga NP2 ni TV-o passive	boy ga young-man ni push-passive	Passive
	(*The boy was pushed by the young man*)	
$\not\Rightarrow H_1$: NP1 ga NP2 o TV-o	男の子 が 若者 を 押した (*The boy pushed the young man*)	ORDER-SUBSET
$\Rightarrow H_2$: NP2 ga NP1 o TV-o	若者 が 男の子 を 押した (*The young man pushed the boy*)	MIXED-SUBSET
	男の子 が カップル を 笑わせている	
P: NP1 ga NP2 o IV causative	boy ga couple o laugh-causative	Causative
	(*The boy is making the couple laugh*)	
$\not\Rightarrow H_1$: NP1 ga IV	男の子 が 笑っている (*The boy is laughing*)	ORDER-SUBSET
$\Rightarrow H_2$: NP2 ga IV	カップル が 笑っている (*The couple is laughing*)	ORDER-SUBSET
	もしかしたら サーファー が 泳いでいる	
P: Factive-adverb NP1 ga IV	perhaps surfer ga swimming	Factive adverb
	(*Perhaps the surfer is swimming*)	
$\not\Rightarrow H_1$: NP1 ga IV	サーファー が 泳いでいる	CONSTITUENT
	(*The surfer is swimming*)	

評価結果（ガーデンパス現象）

Model	Train	Correct: *Entailment*						Correct: *Non-entailment*					
		Normal	Double-o	Punct.	Select.	Wa	Scramb.	Normal	Double-o	Punct.	Select.	Wa	Scramb.
Ja	JSICK	90.2±0.09	90.8±0.10	86.8±0.11	82.9±0.15	84.1±0.13	90.6±0.08	9.3±0.07	11.9±0.11	10.2±0.08	14.1±0.13	13.8±0.11	7.2±0.06
	+JaNLI	99.0±0.00	99.2±0.01	99.4±0.01	98.8±0.01	98.6±0.02	98.7±0.01	91.2±0.13	78.3±0.32	83.0±0.27	87.8±0.19	87.8±0.19	86.9±0.14
	JSNLI	98.3±0.01	95.3±0.03	99.4±0.00	98.8±0.02	99.3±0.00	98.6±0.02	2.0±0.03	3.7±0.04	1.8±0.02	0.6±0.01	2.8±0.03	1.5±0.02
	+JaNLI	83.2±0.07	88.2±0.01	86.5±0.08	92.8±0.09	88.8±0.09	82.8±0.07	58.0±0.16	54.8±0.14	53.1±0.20	49.4±0.19	47.7±0.17	55.9±0.09
Multi	JSICK	62.7±0.55	64.0±0.56	59.8±0.53	62.9±0.55	62.4±0.54	62.5±0.55	35.2±0.56	34.2±0.57	35.8±0.56	35.8±0.56	36.2±0.55	37.9±0.54
	+JaNLI	33.8±0.35	34.8±0.39	30.8±0.28	35.4±0.33	32.4±0.33	27.8±0.32	81.2±0.26	74.9±0.36	84.0±0.19	78.7±0.26	82.8±0.20	80.6±0.24
	JSNLI	98.7±0.01	97.1±0.01	99.6±0.01	99.8±0.00	99.2±0.01	98.7±0.02	0.6±0.01	1.8±0.02	0.2±0.00	0.2±0.00	1.1±0.01	0.8±0.01
	+JaNLI	28.3±0.49	29.8±0.52	30.8±0.53	32.2±0.56	30.9±0.54	29.3±0.51	79.8±0.35	79.2±0.36	78.7±0.37	74.2±0.45	77.9±0.38	78.3±0.38
Human		95.0±0.02	96.7±0.06	100.0±0.00	98.3±0.03	98.3±0.03	97.5±0.03	90.8±0.14	96.7±0.12	91.7±0.10	91.0±0.05	95.0±0.22	96.7±0.04

- 人はガーデンパス文の解釈を簡単にするファクターが含まれているほうが、（わずかであるが）正答率が高い傾向
- モデルはファクターの有無を区別していない傾向

本発表のまとめ

- 深層学習モデルがだまされやすいヒューリスティクスごとに日本語の言語現象に基づく推論テンプレートを設計し、推論データセットを自動構築
- 日本語・多言語BERTが構成的に推論できているのか評価
  →ヒューリスティクスで含意関係を予測し、人にとっては容易な構成的な推論に汎化しない傾向
- 理論言語学に基づくデータセット自動構築は、質の良いデータ拡張手法としても有用な可能性
  - 機能語が重要な役割を果たす言語現象（スクランブリング・受身など）はデータ拡張では捉えるのが困難な可能性

ご清聴ありがとうございました！

JaNLIデータセット：https://github.com/verypluming/JaNLI

谷中　瞳：hyanaka@is.s.u-tokyo.ac.jp