Evaluating NLI Models Using Formal Logic

LING 575C Group 5

Sicong Huang, Shunjie Wang, Yuanbo Xu, Jize Cao

NLI

- Natural language Inference
- Textual entailment

Machines' capability of deep understanding of language that goes beyond what is explicitly expressed, rather relying on new conclusions inferred from knowledge about how the world works. (Bowman, Angeli, Potts, and Manning 2015)

Ex.

“Jack needed some money, so he went and shook his piggy bank. He was disappointed when it made no sound.” (Minsky 2000)

Ex.

“Jack needed some money, so he went and shook his piggy bank. He was disappointed when it made no sound.” (Minsky 2000)

Jack didn’t find any money

Ex.

“Jack needed some money, so he went and shook his piggy bank. He was disappointed when it made no sound.” (Minsky 2000)

He didn’t find any money

- Linguistic knowledge.
- Commonsense knowledge

Why

Besides “deep understanding of language”

Also helpful for

- Question answering
- Information extraction
- Summarization
- Machine translation evaluation
- ...

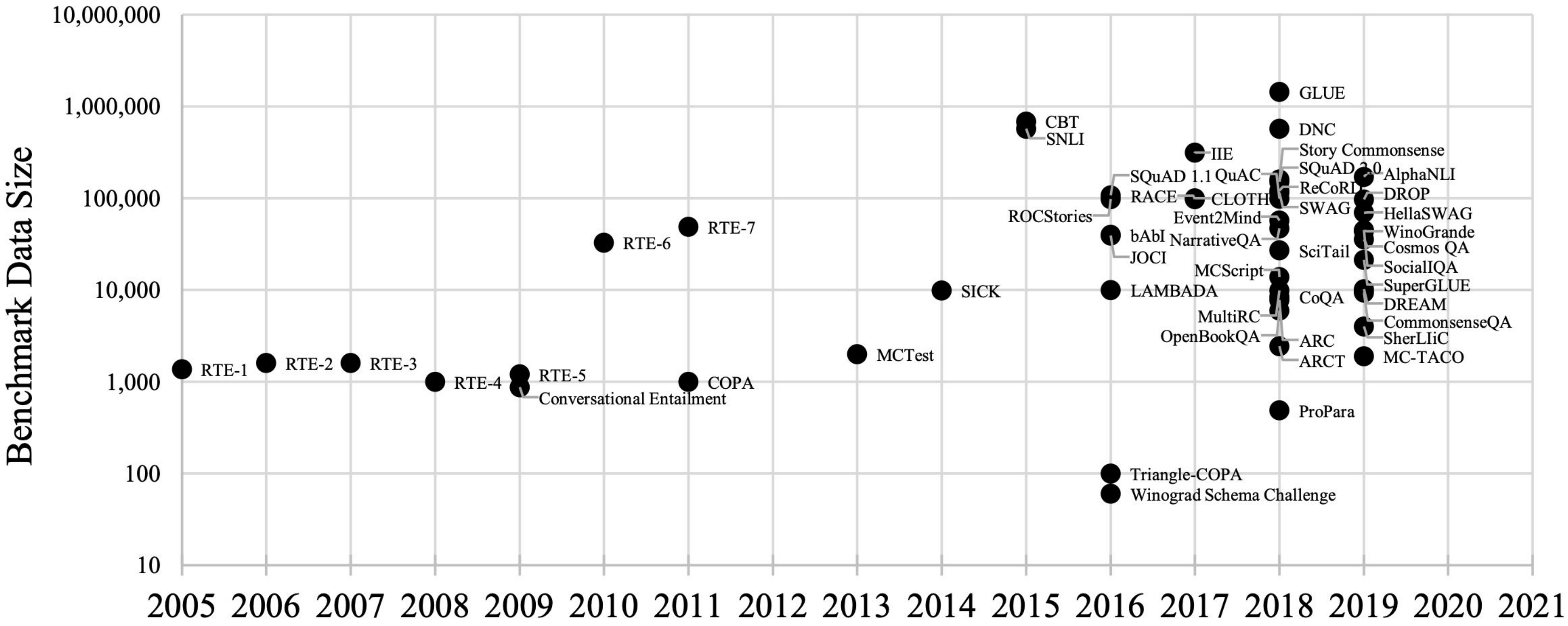
Recognizing Textual Entailment (Dagan et al. 2005)

- Given two text fragments
- T (the entailing text)
- H (the entailed hypothesis)
- T entails H, if a human reading T would infer that H is most likely true

RTE Examples

Text	Hypothesis	Label
Norway's most famous painting, "The Scream" by Edvard Munch, was recovered Saturday, almost three months after it was stolen from an Oslo museum.	Edvard Munch painted "The Scream".	True
Bush returned to the White House late Saturday while his running mate was off campaigning in the West.	Bush left the White House	False

Growing Number of Datasets



Shorks, Gao, and Chai (2019)

Motivation

#	Attacked Sentence
1	Mr. Tsai is a very **orig~~i~~nal** artist in his medium, and what time is it there?
2	Old-form moviemaking at its **be~~s~~t**.
3	My reaction in a word: **disapponitment**.
4	a **painfulily funtny** ode to **gbad** behavior.

Learning to Discriminate Perturbations for Blocking Adversarial Attacks in Text Classification
(Zhou et. al 2019)

Motivation

Original Text Prediction: **Entailment** (Confidence = 86%)
Premise: *A runner wearing purple strives for the finish line.* **Hypothesis:** *A runner wants to head for the finish line.*

Adversarial Text Prediction: **Contradiction** (Confidence = 43%)
Premise: *A runner wearing purple strives for the finish line.* **Hypothesis:** *A **racer** wants to head for the finish line.*

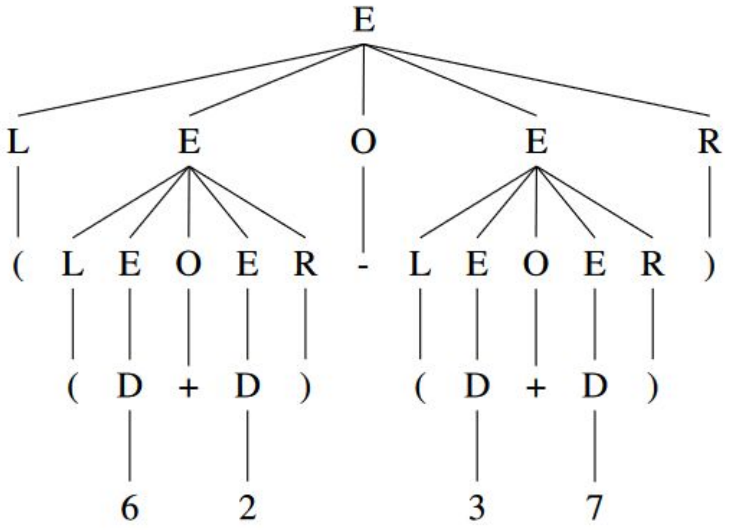
Generating Natural Language Adversarial Examples (Alzantot et. al 2018)

P: The dog did not eat all of the chickens. **H**: The dog ate all of the chickens. **S**: entails (score 56.5%)
P: The red box is in the blue box. **H**: The blue box is in the red box. **S**: entails (score 92.1%)

AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples (Kang et. al 2018)

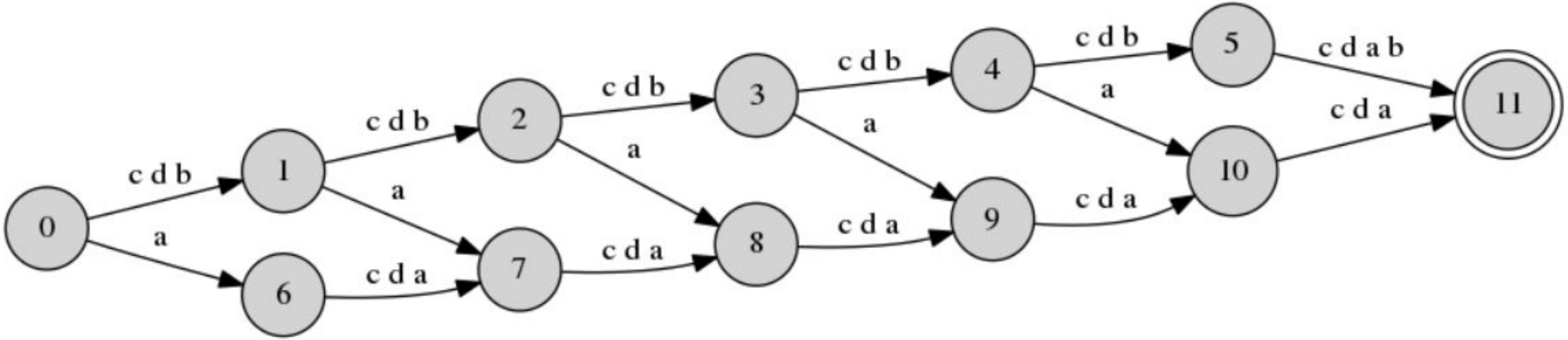
Formal Languages in NN Evaluation

Syntax	Meaning
$E \rightarrow L\ E_1\ O\ E_2\ R$	$[E] = [O]([E_1], [E_2])$
$E \rightarrow D$	$[E] = [D]$
$O \rightarrow +$	$[O] = \lambda x, y.x + y \bmod 10$
$O \rightarrow -$	$[O] = \lambda x, y.x - y \bmod 10$
$L \rightarrow ($	
$R \rightarrow)$	
$D \rightarrow 0$	$[D] = 0$
$\vdots$	$\vdots$
$D \rightarrow 9$	$[D] = 9$



Correlating Neural and Symbolic Representations of Language (Chrupała, Alishahi 2019)

Formal Languages in NN Evaluation

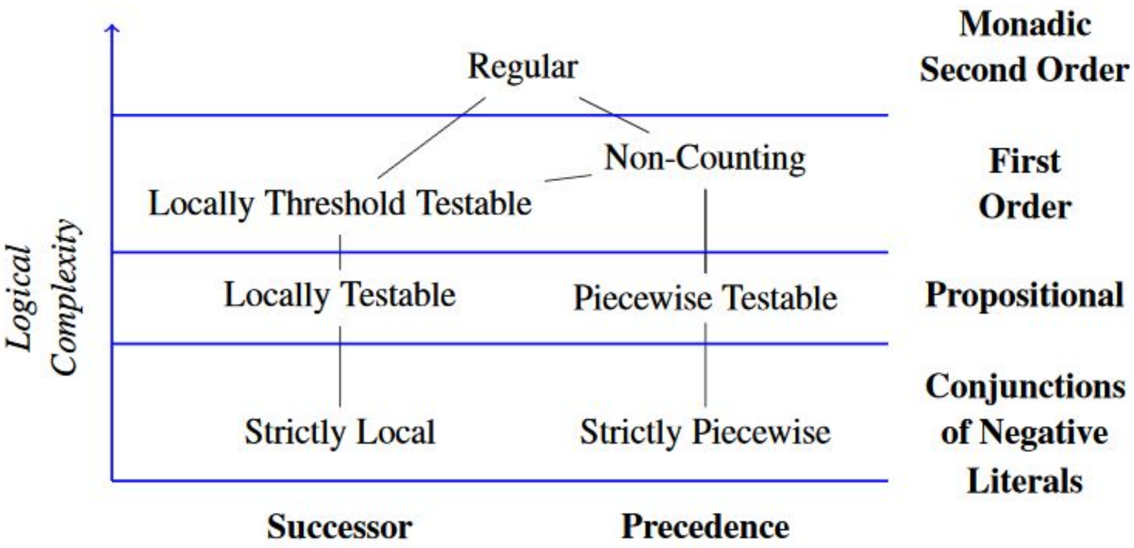


$$\Sigma = \{a, b, c, d\}$$

$$G_{SP2} = \{aa,\ ac,\ ad,\ ba,\ bb,\ bc,\ bd,\ ca,\ cb,\ cc,\ cd,\ da,\ db,\ dc,\ dd\}$$

Using Regular Languages to Explore the Representational Capacity of Recurrent Neural Architectures (Mahalunkar & Kelleher 2018)

Formal Languages in NN Evaluation



Subregular Complexity and Deep Learning (Avcu et al. 2017)

Propositional Logic

If S is a sentence, $\neg S$ is a sentence (negation)

If S_1 and S_2 are sentences, $S_1 \wedge S_2$ is a sentence (conjunction)

If S_1 and S_2 are sentences, $S_1 \vee S_2$ is a sentence (disjunction)

If S_1 and S_2 are sentences, $S_1 \Rightarrow S_2$ is a sentence (implication)

If S_1 and S_2 are sentences, $S_1 \Leftrightarrow S_2$ is a sentence (biconditional)

Courtesy: AIMA Slides, Russell and Norvig

Propositional Logic

$\neg S$	is true iff	S	is false		
$S_1 \wedge S_2$	is true iff	S_1	is true **and**	S_2	is true
$S_1 \vee S_2$	is true iff	S_1	is true **or**	S_2	is true
$S_1 \;\Rightarrow\; S_2$	is true iff	S_1	is false **or**	S_2	is true
i.e.,	is false iff	S_1	is true **and**	S_2	is false
$S_1 \;\Leftrightarrow\; S_2$	is true iff	$S_1 \;\Rightarrow\; S_2$	is true **and**	$S_2 \;\Rightarrow\; S_1$	is true

Courtesy: AIMA Slides, Russell and Norvig

Propositional Logic

P	Q	$\neg P$	$P \wedge Q$	$P \vee Q$	$P \Rightarrow Q$	$P \Leftrightarrow Q$
false	*false*	*true*	*false*	*false*	*true*	*true*
false	*true*	*true*	*false*	*true*	*true*	*false*
true	*false*	*false*	*false*	*true*	*false*	*false*
true	*true*	*false*	*true*	*true*	*true*	*true*

Courtesy: AIMA Slides, Russell and Norvig

First-Order Logic

Constants	$KingJohn,\ 2,\ UCB, \ldots$
Predicates	$Brother,\ >, \ldots$
Functions	$Sqrt,\ LeftLegOf, \ldots$
Variables	$x,\ y,\ a,\ b, \ldots$
Connectives	$\wedge\ \vee\ \neg\ \Rightarrow\ \Leftrightarrow$
Equality	$=$
Quantifiers	$\forall\ \exists$

Courtesy: AIMA Slides, Russell and Norvig

First-Order Logic: Atomic Sentences

$$Brother(KingJohn, RichardTheLionheart)$$

$$> (Length(LeftLegOf(Richard)), Length(LeftLegOf(KingJohn)))$$

Courtesy: AIMA Slides, Russell and Norvig

First-Order Logic: Complex Sentences

$$Sibling(KingJohn, Richard) \Rightarrow Sibling(Richard, KingJohn)$$
$$>(1,2) \vee \leq(1,2)$$
$$>(1,2) \wedge \neg{>}(1,2)$$

Courtesy: AIMA Slides, Russell and Norvig

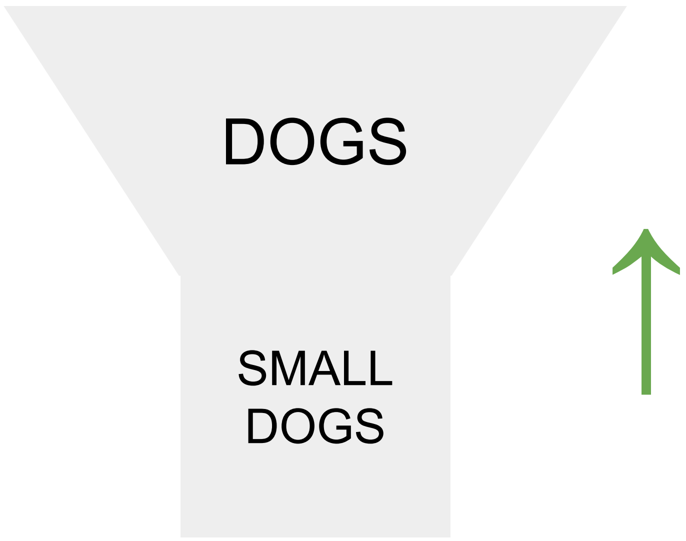
First-Order Logic: Universal and Existential Quantifiers

$$\forall x \;\; At(x, Berkeley) \;\Rightarrow\; Smart(x)$$

$$\exists x \;\; At(x, Stanford) \land Smart(x)$$

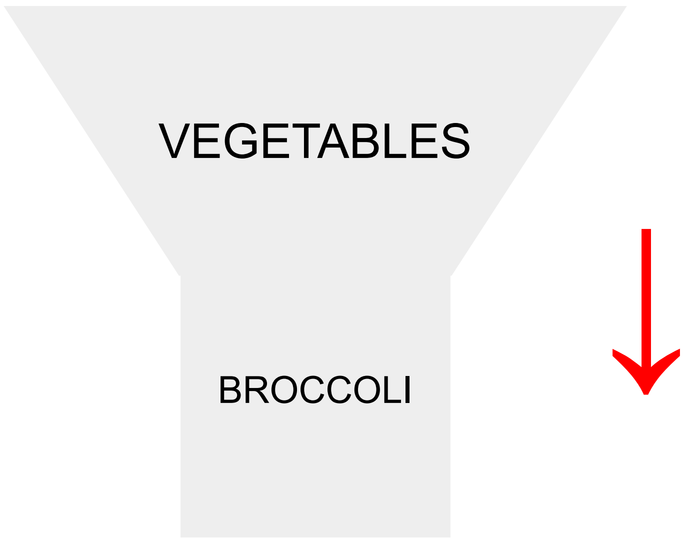
Courtesy: AIMA Slides, Russell and Norvig

Monotonicity: Upward Entailing



The cat chased a dog ⇐ The cat chased a small dog

Monotonicity: Downward Entailing



I don’t like vegetables ⇒ I don’t like broccoli

Natural Logic Quantifiers

some$^{\uparrow\uparrow}$
Some mammals fly
$\Rightarrow$ *Some animals fly*
$\Rightarrow$ *Some mammals move*

all$^{\downarrow\uparrow}$
All ducks fly
$\Rightarrow$ *All mallards fly*
$\Rightarrow$ *All ducks move*

no$^{\downarrow\downarrow}$
No dogs fly
$\Rightarrow$ *No poodles fly*
$\Rightarrow$ *No dogs hover*

not every$^{\uparrow\downarrow}$
Not every bird flies
$\Rightarrow$ *Not every animal flies*
$\Rightarrow$ *Not every bird hovers*

Courtesy: Natural Logic for Textual Inference (MacCartney and Manning 2007)

Compositionality

	↓M	↑M	=M
↑M	↑M	↓M	=M
↓M	↓M	↑M	=M
=M	=M	=M	=M

Courtesy: Natural Logic for Textual Inference (MacCartney and Manning 2007)

A logical-based corpus for cross-lingual evaluation*

Felipe Salvatore[1], **Marcelo Finger**[1†] and **R. Hirata Jr**[1‡]
[1]Department of Computer Science, Instituto de Matemática e Estatística,
University of São Paulo, Brazil
{felsal, mfinger, hirata}@ime.usp.br

Lexical inference vs Structural inference

1. “*A woman plays with my dog*”

   “*A person plays with my dog*”

2. “*Jenny and Sally play with my dog*”

   “*Jenny plays with my dog*”

Template Language

- A formal Language to generate instances
- Two basic entities:
  - People (*Pe*)
  - Place(*Pl*)
- Three basic relation:
  - V(x,y): x has visited y
  - x>y: x is taller than y
  - x=y: x is as tall as y
- Realisation
  - $r(Pe) \cap r(Pl) = \varnothing$
  - $r\_train(Pe) \cap r\_test(Pe) = \varnothing$
  - $r\_train(Pl) \cap r\_test(Pl) = \varnothing$

Data Generation in 7 tasks

- Simple Negation
  - P: $\{V(x1, p1), V(x2, p2)\}$ "Charles has visited Chile, Jos has visited Japan"
  - H1: $\neg V(x2, p2)$ "Charles didn't visit Japan"
  - H2: $\neg V(x, p)$ "Lana didn't visit France"
- Boolean Coordination
  - P: $\{V(x1, p) \wedge V(x2, p) \wedge V(x3, p)\}$ "Felix, Ronnie, and Tyler have visited Bolivia"
  - H1: $\neg V(x3, p)$ "Tyler didn't visit Bolivia"
  - H2: $\neg V(x, p)$ "Bruce didn't visit Bolivia"
- Qualification
  - P: $\{\forall x \forall p V(x, p)\}$ "Everyone has visited everyplace"
  - H1: $\neg V(x, p)$ "Timothy didn't visit El Salvador"
  - H2: $\neg V(x, x1)$ "Timothy didn't visit Anthony"

Data Generation in 7 tasks

- Definite Description
  - P: $\{x1 = \iota y \forall p V(y, p), V(x1, x2)\}$ "Carlos is the person that has visited every place, Carlos has visited John"
  - H1: $\neg V(x1, p)$ "Carlos did not visit Germany"
  - H2: $\neg V(x2, p)$ "John did not visit Germany"
- Comparatives
  - P: $\{x1 > x2, x2 > x3\}$ "Francis is taller than Joe, Joe is taller than Ryan"
  - H1: $x1 > x3$ "Francis is taller than Ryan"
  - H2: $x3 > x1$ "Ryan is taller than Francis"
- Counting
  - P: $\{\exists{=3}p V(x1, p) \land \exists{=2}x V(x1, x)\}$ "Philip has visited only three places and only two people"
  - H1: $V(x1, x2)$ "Philip has visited John"
  - H2: $V(x1, x2) \land V(x1, x3) \land V(x1, x4)$ "Philip has visited John, Carla, and Bruce"
- Mixed : Combination of all 6 tasks above

Dataset statistics

Task	Vocab size	Vocab inter-section	Mean input length	Max input length
1 (Eng)	3561	77	230.6	459
2 (Eng)	4117	128	151.4	343
3 (Eng)	3117	70	101.5	329
4 (Eng)	1878	62	100.81	134
5 (Eng)	1311	25	208.8	377
6 (Eng)	3900	150	168.4	468
7 (Eng)	3775	162	160.6	466
1 (Pt)	7762	254	209.4	445
2 (Pt)	9990	393	148.5	388
3 (Pt)	5930	212	102.7	395
4 (Pt)	5540	135	91.8	140
5 (Pt)	5970	114	235.2	462
6 (Pt)	9535	386	87.8	531
7 (Pt)	8880	391	159.9	487

Model

- Baseline: Random Forest with BOW input
- RNN
- GRU
- LSTM
- Bert_eng, Bert_mult, Bert_chi

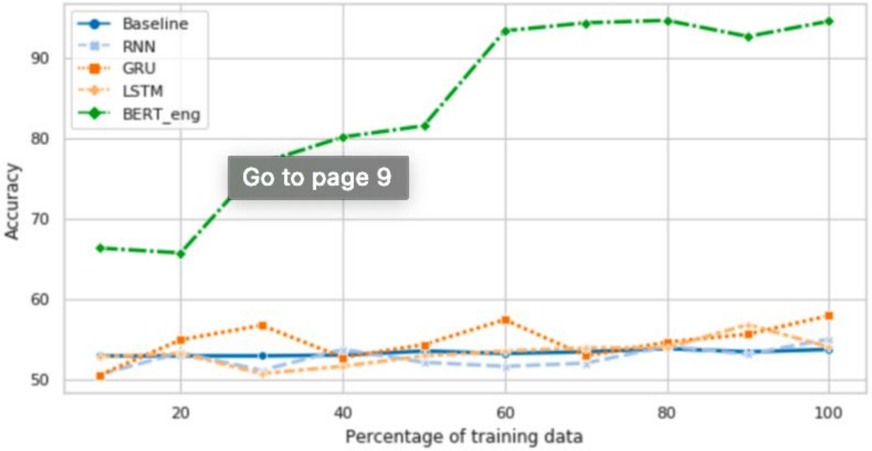
Experimental Setting

Questions	**Experimental setting**
How the different models perform on the proposed tasks?	$r\_train(Pe) \cap r\_test(Pe) = \varnothing$ $r\_train(Pl) \cap r\_test(Pl) = \varnothing$
How much each model rely on the occurrence of non-logical words?	$r\_train(Pe) = r\_test(Pe)$ $r\_train(Pl) = r\_test(Pl)$
Can cross-lingual transfer learning be successfully used for the Portuguese realization of those tasks?	BERT_eng, BERT_mult, BERT_chi.
Is the dataset biased? Are the models learning some unexpected text pattern?	Noise label Premise only Hypothesis only

Result

- How the different models perform on the proposed tasks?

Task	Base	RNN	GRU	LSTM	BERT
1 (Eng)	52.1	50.1	50.6	50.4	**99.8**
2 (Eng)	50.7	50.2	50.2	50.8	**100**
3 (Eng)	63.5	50.3	66.1	63.5	**90.5**
4 (Eng)	51.0	51.7	52.7	51.6	**100**
5 (Eng)	50.6	50.1	50.2	50.2	**100**
6 (Eng)	55.5	84.4	82.7	75.1	**87.5**
7 (Eng)	54.1	50.9	53.7	50.0	**94.6**
Avg.	53.9	55.4	58.0	56.2	**96.1**
1 (Pt)	53.9	50.1	50.2	50.0	**99.9**
2 (Pt)	49.8	50.0	50.0	50.0	**99.9**
3 (Pt)	61.7	50.0	70.6	50.1	**78.7**
4 (Pt)	50.9	50.0	50.4	50.0	**100**
5 (Pt)	49.9	50.1	50.8	50.0	**99.8**
6 (Pt)	58.9	66.4	**79.7**	67.2	79.1
7 (Pt)	55.4	51.1	51.6	51.1	**82.7**
Avg.	54.4	52.6	57.6	52.6	**91.4**



Go to page 9

Result

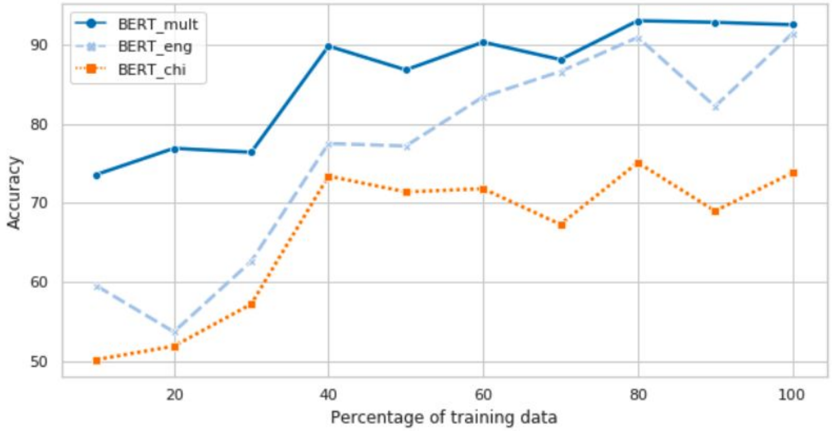
- How much each model rely on the occurrence of non-logical words?

Model	Avg. improvement using (ii)
Baseline	17.6%
GRU	9.6%
BERT_eng	5.3%
LSTM	4.25%
RNN	1.3%

Recurrent models are relying more on noun phrases than Bert*

Result

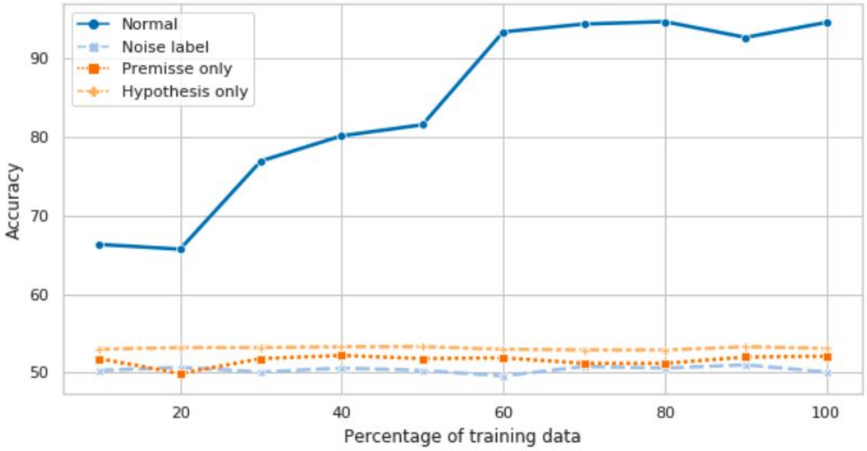
- Can cross-lingual transfer learning be successfully used for the Portuguese realization of those tasks?



Multilingual Bert perform better

Result

- Is the dataset biased? Are the models learning some unexpected text pattern?



Bert_eng is not memorizing textual pattern

Probing Natural Language Inference Models through Semantic Fragments

Kyle Richardson[†] and **Hai Hu**[‡] and **Lawrence S. Moss**[‡] and **Ashish Sabharwal**[†]

[†]Allen Institute for AI, Seattle, WA, USA
[‡]Indiana University, Bloomington, IN, USA
[†]{kyler,ashishs}@allenai.org, [‡]{huhai, lmoss}@indiana.edu

- Follow-up work on previous paper (December 2019)
- 2-value-logic -> 3-value-logic
- Monotonicity fragment

2-value-logic -> 3-value-logic

- ENTAILMENT: A -> B
- CONTRADICTION: A -> ¬B
- NEUTRAL: Neither entailment of contradiction

Template generation

Logic Fragment	Rule Template: [premise], { hypothesis$_1$, ... } ⇒ label; Labeled Examples (simplified)		
Negation	[only-did-p(x)], ¬p(x)	⇒ CONTRADICTION	Dave$_x$ has only visited Israel$_p$, Dave$_x$ didn't$_\neg$ visit Israel$_p$
	[only-did-p(x)], ¬p$'(x)$	⇒ ENTAILMENT	Dave$_x$ has only visited Israel$_p$, Dave$_x$ didn't$_\neg$ visit Russia$_{p'}$
	[only-did-p(x)], ¬p(x')	⇒ NEUTRAL	Dave$_x$ has only visited Israel$_p$, Bill$_x$ didn't$_\neg$ visit Israel$_p$
Boolean	[p(x_1) ∧ ...∧p(x_n)], ¬p(x_j)	⇒ CONTRADICTION	Dustin$_{x_1}$, Milton$_{x_2}$, ... have only visited Equador$_p$; Dustinx_1 didn't$_\neg$ visit Equador$_p$
	[p$_1(x_1)$ ∧ ...∧p$_n(x_n)$], ¬p$_j(x')$	⇒ NEUTRAL	Dustin$_x$ only visited$_p$ Portugal$_1$ and Spain$_2$; James$_{x'}$ didn't$_\neg$ visit$_p$ Spain$'$
	[p$_1(x)$ ∧ ...∧p$_n(x)$], ¬p$'(x)$	⇒ ENTAILMENT	Dustin$_x$ only visited$_p$ Portugal$_1$ and Spain$_2$; Dustin$_x$ didn't$_\neg$ visit$_p$ Germany$'$
Conditional	[(p → q) ∧ p], q	⇒ ENTAILMENT	Dave visited Israel$_p$ and if Dave visited Israel$_p$ then$_\rightarrow$ Bill visited Russia$_q$; Bill visited Russia$_q$.
	[(p → q) ∧ p], ¬q	⇒ CONTRADICTION	Dave visited Israel$_p$ and if Dave visited Israel$_p$ then$_\rightarrow$ Bill visited Russia$_q$; Bill didn't visit Russia$_p$.
	[(p → q) ∧ ¬p], {q,¬q}	⇒ NEUTRAL	Dave didn't visit Israel$_p$, and if Dave visited Israel$_p$ then$_\rightarrow$ Bill visited Russia$_q$; Bill visited Russia$_p$.
Quantifier	[$\forall x.\forall y.$ p(x,y)], $\exists x.\iota y.$ ¬p(x,y)	⇒ CONTRADICTION	Everyone$_{\forall x}$ visited$_p$ every$_\forall$ country$_y$; Someone$_{\exists x}$ didn't$_\neg$ visit$_p$ Jordan$_{\iota y}$
	[$\exists x.\forall y.$ p(x,y)], $\iota x.\exists y.$ {¬p(x,y),p(x,y)}	⇒ NEUTRAL	Someone$_{\exists x}$ visited$_p$ every$_\forall$ person$_y$; Tim$_{\iota x}$ didn't$_\neg$ visit$_p$ someone$_{\exists y}$
	[$\exists x.\forall y.$ p(x,y)], $\exists x.\iota y.$ p(x,y)	⇒ ENTAILMENT	Someone$_{\exists x}$ visited$_p$ every$_\forall$ person$_y$; A person$_{\exists x}$ visited$_p$ Mark$_{\iota y}$

2-value-logic -> 3-value-logic

Monotonicity Fragments

Upward entailing/monotone tokens: Entail somethings “greater or equal to them”

Downward entailing/monotone tokens: Entail somethings “less than them”

Premise: All$_\uparrow$ dogs$_\downarrow$ chased$_\uparrow$ some$_\uparrow$ cat$_\uparrow$
Hypothesis 1: All small dogs chased a cat. (ENTAILMENT)
Hypothesis 2: All mammals chased a cat. (NEUTRAL)
Hypothesis 3: ALL dogs don’t chased some cat (CONTRADICTION)

(Generate fragment through substitution)

Monotonicity Fragments

premise: All$^{\uparrow}$ black$^{\downarrow}$ mammals$^{\downarrow}$ saw$^{\uparrow}$ exactly$^{=}$ 5$^{=}$ stallions$^{=}$ who$^{=}$ danced$^{=}$

Some$^{\uparrow}$ black$^{\uparrow}$ mammal$^{\uparrow}$ saw$^{\uparrow}$ exactly$^{=}$ 5$^{=}$ stallions$^{=}$ who$^{=}$ danced$^{=}$

All$^{\uparrow}$ black$^{\downarrow}$ dogs$^{\downarrow}$ saw$^{\uparrow}$ exactly$^{=}$ 5$^{=}$ stallions$^{=}$ who$^{=}$ danced$^{=}$

Some$^{\uparrow}$ mammal$^{\uparrow}$ saw$^{\uparrow}$ exactly$^{=}$ 5$^{=}$ stallions$^{=}$ who$^{=}$ danced$^{=}$

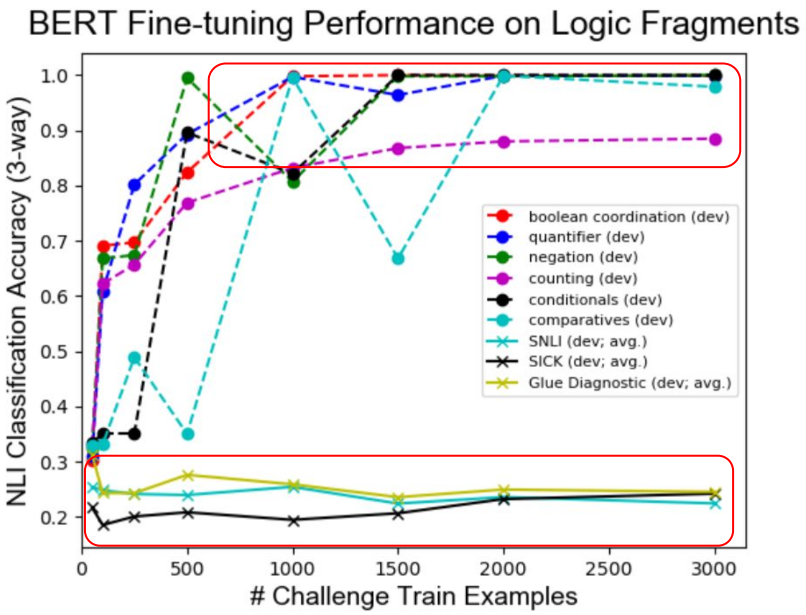
All$^{\uparrow}$ black$^{\downarrow}$ doodles$^{\downarrow}$ saw$^{\uparrow}$ exactly$^{=}$ 5$^{=}$ stallions$^{=}$ who$^{=}$ danced$^{=}$

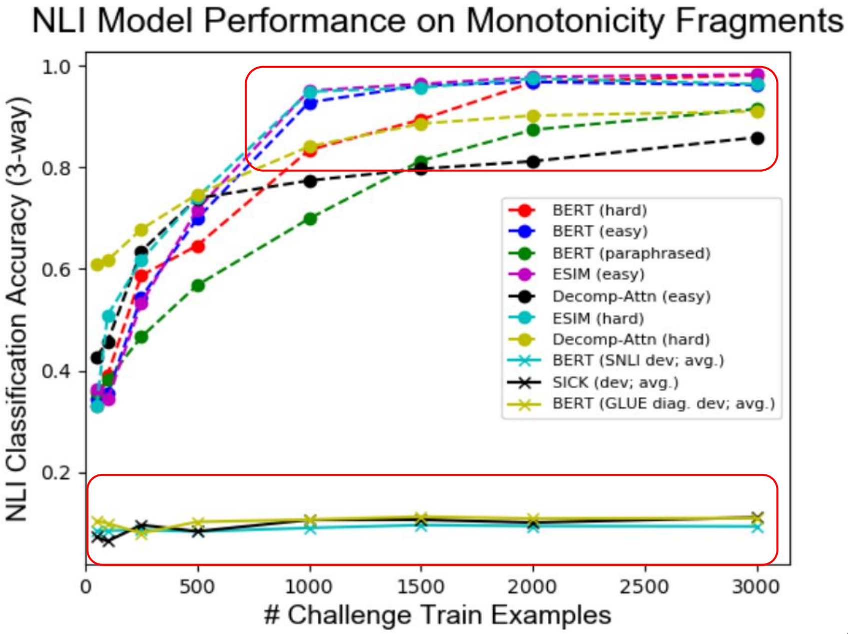
Experimental Setting

Questions	**Experimental setting**
Is this particular fragment learnable from scratch using existing NLI architectures？	BERT … NLI models
How well do large state-of-the-art pre-trained NLI models？	BERT_SNLI BERT_SNLI+MNLI … NLI models
Can existing models be quickly re-trained or re-purposed to be robust on these fragments？	Re-finetune models on both the original dataset and the challenge datasets

Result

- Is this particular fragment learnable from scratch using existing NLI architectures (not pretrain on any NLI datasets)？





- How well do large state-of-the-art pre-trained NLI models？

$\text{Model}_{\text{train\_data}}$	SNLI Test		Logic Fragments (Avg. of 6)	Mono. Fragments (Avg. over 2)	Breaking NLI
			Random/Trained Baselines		
Majority Baseline	34.2		34.6	34.0	–
Hypothesis-Only biLSTM	69.0		49.3	56.7	–
Premise-Only biLSTM	–		44.3	57.4	–
Premise+Hyp. biLSTM	–		52.0	59.1	–
			Pre-Trained NLI Models		
$\textbf{BERT}_{\text{SNLI+MNLI}}$	91.0		47.3	62.8	95.8
$\textbf{BERT}_{\text{SNLI}}$	90.7		46.1	56.8	94.3
$\textbf{Decomp-Attn}_{\text{SNLI}}$	86.4		42.1	48.4	49.9
$\textbf{ESIM}_{\text{SNLI}}$	88.5		44.3	62.8	68.7
	MNLI Dev (Avg.)			**Re-Trained Models with Fragments (frag)**	
$\textbf{BERT}_{\text{SNLI+MNLI+frag}}$	83.7	(↓ 1.3)	98.0	97.8	-
$\textbf{ESIM}_{\text{MNLI+frag}}$	72.0	(↓ 5.9)	86.4	96.5	-
$\textbf{Decomp-Attn}_{\text{MNLI+frag}}$	66.1	(↓ 6.7)	71.7	93.5	-

- Can existing models be quickly re-trained or re-purposed to be robust on these fragments？