

Natural Language Inference



Outline

- The task definition
- Datasets and models
- Examples

Natural language understanding

Suppose someone claims that a program can understand natural language, how can we test for that?

Natural language understanding

Suppose someone claims that a program can understand natural language, how can we test for that?

How might you conduct this test?

Natural language understanding

Suppose someone claims that a program can understand natural language, how can we test for that?

Some ideas:

- Play the imitation game
- Have it answer questions
- See if it makes the same kind of inferences as people

Natural language understanding

Suppose someone claims that a program can understand natural language, how can we test for that?

Some ideas:

- Play the imitation game
- Have it answer questions
- See if it makes the same kind of inferences as people

The tricky part: **How do we conduct these tests without having a human in the loop?**

One answer: Recognizing textual entailment

Premise Before it moved to Chicago, aerospace manufacturer Boeing was the largest company in Seattle.

Hypothesis Boeing is a Chicago-based aerospace manufacturer.

Given a premise and a hypothesis (both in natural language),

- Would a person who reads the premise say that the hypothesis is true?
- Would a person who reads the premise say that the hypothesis is false?
- Or neither?

One answer: Recognizing textual entailment

Premise Before it moved to Chicago, aerospace manufacturer Boeing was the largest company in Seattle.

Hypothesis Boeing is a Chicago-based aerospace manufacturer.

Given a premise and a hypothesis (both in natural language),

- Would a person who reads the premise say that the hypothesis is true?
`Entail`
- Would a person who reads the premise say that the hypothesis is false?
`Contradict`
- Or neither?
`Neutral`

One answer: Recognizing textual entailment

Premise Before it moved to Chicago, aerospace manufacturer Boeing was the largest company in Seattle.

What is the label for this pair?

Hypothesis Boeing is a Chicago-based aerospace manufacturer.

Given a premise and a hypothesis (both in natural language),

- Would a person who reads the premise say that the hypothesis is true?
Entail
- Would a person who reads the premise say that the hypothesis is false?
Contradict
- Or neither?
Neutral

One answer: Recognizing textual entailment

Premise Before it moved to Chicago, aerospace manufacturer

Importantly, we can label datasets for this three-class classification task

And we can build models and evaluate them

What is the label
for this pair?

Given a premise and a hypothesis (both in natural language),

- Would a person who reads the premise say that the hypothesis is true?

Entail

- Would a person who reads the premise say that the hypothesis is false?

Contradict

- Or neither?

Neutral

Outline

- The task definition
- Datasets and models
- Examples

The Recognizing Textual Entailment challenge

A series of annual challenge tasks

Textual entailment is defined as a directional relationship between pairs of text expressions, denoted by T (the entailing “*Text*”) and H (the entailed “*Hypothesis*”). We say that T entails H if humans reading T would typically infer that H is most likely true.

The Recognizing Textual Entailment challenge

A series of annual challenge tasks

Textual entailment is defined as a **directional relationship** between pairs of text expressions, denoted by T (the entailing “*Text*”) and H (the entailed “*Hypothesis*”). We say that T entails H if humans reading T would typically infer that H is most likely true.

The Recognizing Textual Entailment challenge

A series of annual challenge tasks

Textual entailment is defined as a directional relationship between pairs of text expressions, denoted by T (the entailing “*Text*”) and H (the entailed “*Hypothesis*”). We say that T entails H if humans reading T would typically infer that H is most likely true.

The Recognizing Textual Entailment challenge

A series of annual challenge tasks

Textual entailment is defined as a directional relationship between pairs of text expressions, denoted by T (the entailing “*Text*”) and H (the entailed “*Hypothesis*”). We say that T entails H if humans reading T would typically infer that H is most likely true.

The Hypothesis H of an entailment pair contradicts the Text T if a human reader would say that H is highly unlikely to be true given the information described in T .

Some representative examples from the RTE task

Text: The purchase of Houston-based LexCorp by BMI for \$2Bn prompted widespread sell-offs by traders as they sought to minimize exposure. LexCorp had been an employee-owned concern since 2008.

Hyp 1: BMI acquired an American company.

Hyp 2: BMI bought employee-owned LexCorp for \$3.4Bn.

Hyp 3: BMI is an employee-owned concern.

Why is entailment interesting?

To be able to correctly assess the entailment relationship between sentences, we need to be able to understand many different linguistic phenomena and perform reasoning with background knowledge

Why is entailment interesting?

To be able to correctly assess the entailment relationship between sentences, we need to be able to understand many different linguistic phenomena and perform reasoning with background knowledge

Task	NLI framing
Paraphrase	text \equiv paraphrase
Summarization	text \sqsupset summary
Information retrieval	query \sqsupset document
Question answering	question \sqsupset answer <i>Who left? \Rightarrow Someone left</i> <i>Someone left \sqsupset Sandy left</i>

Why is entailment interesting?

To be able to correctly assess the entailment relationship between sentences, we need to be able to understand many different linguistic phenomena and perform reasoning with background knowledge

Phenomenon	Description	Example(s)
Hypernymy	“IsA”	“Honda Civic” \models “car”
Synonymy	“is often interchangeable with”	“beast” \models “animal”
Metonymy	“can be used to represent”	“wheels” for “car”; “the suits at the bank awarded themselves a pay raise”
Antonymy	“opposite”	“rise” and “fall”
Scalar implicature	relative proportions and quantities	“3 in 10 doctors” \models “some doctors” but not “most doctors”
Thematic Roles	for a given predicate, “who did what to whom”	“John broke the vase” \models “The vase was broken by John”

Why is entailment interesting?

To be able to correctly assess the entailment relationship between sentences, we need to be able to understand many different linguistic phenomena and perform reasoning with background knowledge

Phenomenon	Example
Implicit argument	“Xavier arrived in America in 1932. Within a year he was running a fashionable hardware store [in America].”
Implicit predicate (e.g., comparisons)	“More people arrived than we thought [would arrive]”
Redundant head	“The fumigation process” \models “The fumigation”.
Implicit argument	“Mary arrived at the party at 8pm. John came [to the party] later.”
Head drop	“Three [people] died and two [people] were injured in a car accident on the M1 today.”
Noun compounds (implicit relations)	“Microcomp CEO Jeff Burns” \models “Jeff Burns is the CEO of Microcomp”; “oatmeal cookie” \models “cookie containing oatmeal”
Possessives	“Einstein’s Theory of Relativity” \models “Einstein discovered the Theory of Relativity”; “Michelangelo’s <i>David</i> ” \models “Michelangelo created the artwork ‘David’”
Implicit quantifiers	“[all] Government employees must submit expense reports.”

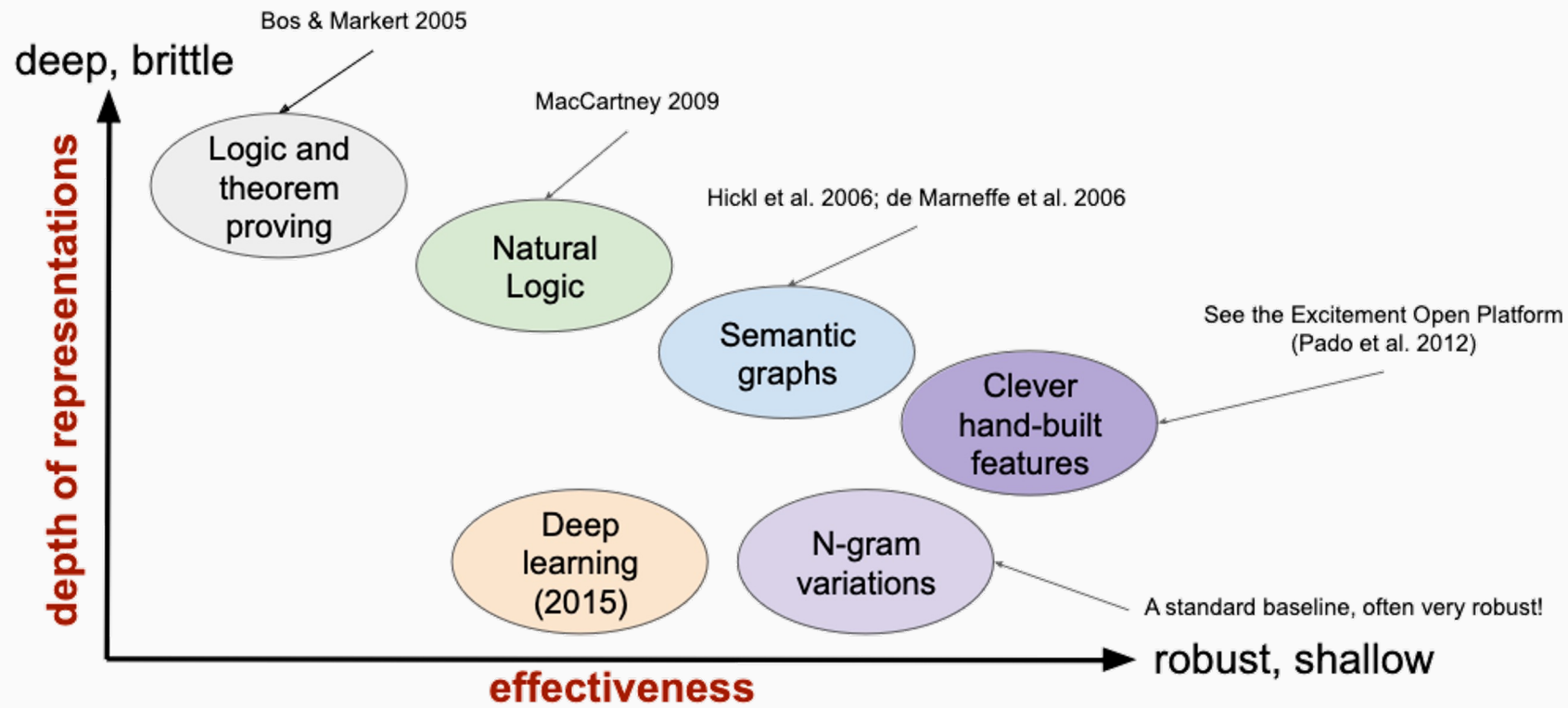
Why is entailment interesting?

To be able to correctly assess the entailment relationship between sentences, we need to be able to understand many different linguistic phenomena and perform reasoning with background knowledge

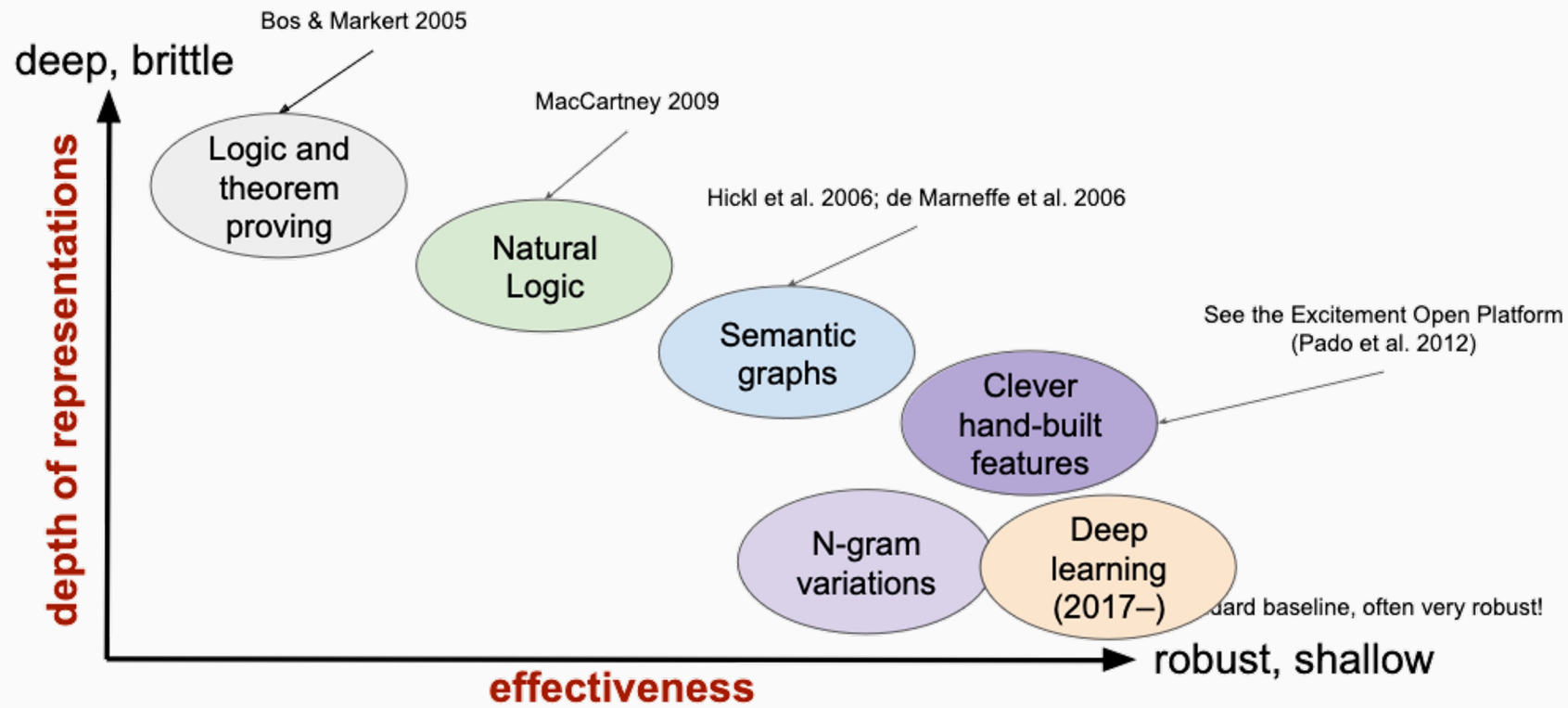
Phenomenon	Example
Implicit argument	“Xavier arrived in America in 1932. Within a year he was running a fashionable hardware store [in America].”
Implicit predicate (e.g., comparisons)	“More people arrived than we thought [would arrive]”
Redundant head	“The fumigation process” \models “The fumigation”.
Implicit argument	“Mary arrived at the party at 8pm. John came [to the party] later.”
Head drop	“Three [people] died and two [people] were injured in a car accident on the M1 today.”
Noun compounds (implicit relations)	“Microcomp CEO Jeff Burns” \models “Jeff Burns is the CEO of Microcomp”; “oatmeal cookie” \models “cookie containing oatmeal”
Possessives	“Einstein’s Theory of Relativity” \models “Einstein discovered the Theory of Relativity”; “Michelangelo’s <i>David</i> ” \models “Michelangelo created the artwork ‘David’”
Implicit quantifiers	“[all] Government employees must submit expense reports.”

And many many more

Models for predicting entailments



Models for predicting entailments



Outline

- The task definition
- Datasets and models
- Examples

The Stanford Natural Language Inference (SNLI) dataset

Bowman et al 2015

- A large crowdsourced dataset
 - All the premises are image captions from the Flickr30K corpus (Young et al. 2014).
 - All the hypotheses were written by crowdworkers.
- Dataset statistics
 - 550,152 train examples, 10K each in dev and test sets
 - Average number of tokens:
 - Premise: 14.1
 - Hypothesis: 8.3
 - Vocabulary: 37,026 words
- 56,951 examples validated by four additional annotators.
 - 58.3% examples with unanimous gold label
 - 91.2% of gold labels match the author's label
 - 0.70 overall Fleiss kappa

Some of the sentences reflect societal stereotypes (Rudinger et al. 2017), which could be problematic

Crowdsourcing approach for SNLI

Instructions

The [Stanford University NLP Group](#) is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely a true** description of the photo.
- Write one alternate caption that **might be a true** description of the photo.
- Write one alternate caption that is **definitely an false** description of the photo.

Photo caption **A little boy in an apron helps his mother cook.**

Definitely correct Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

Write a sentence that follows from the given caption.

Maybe correct Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

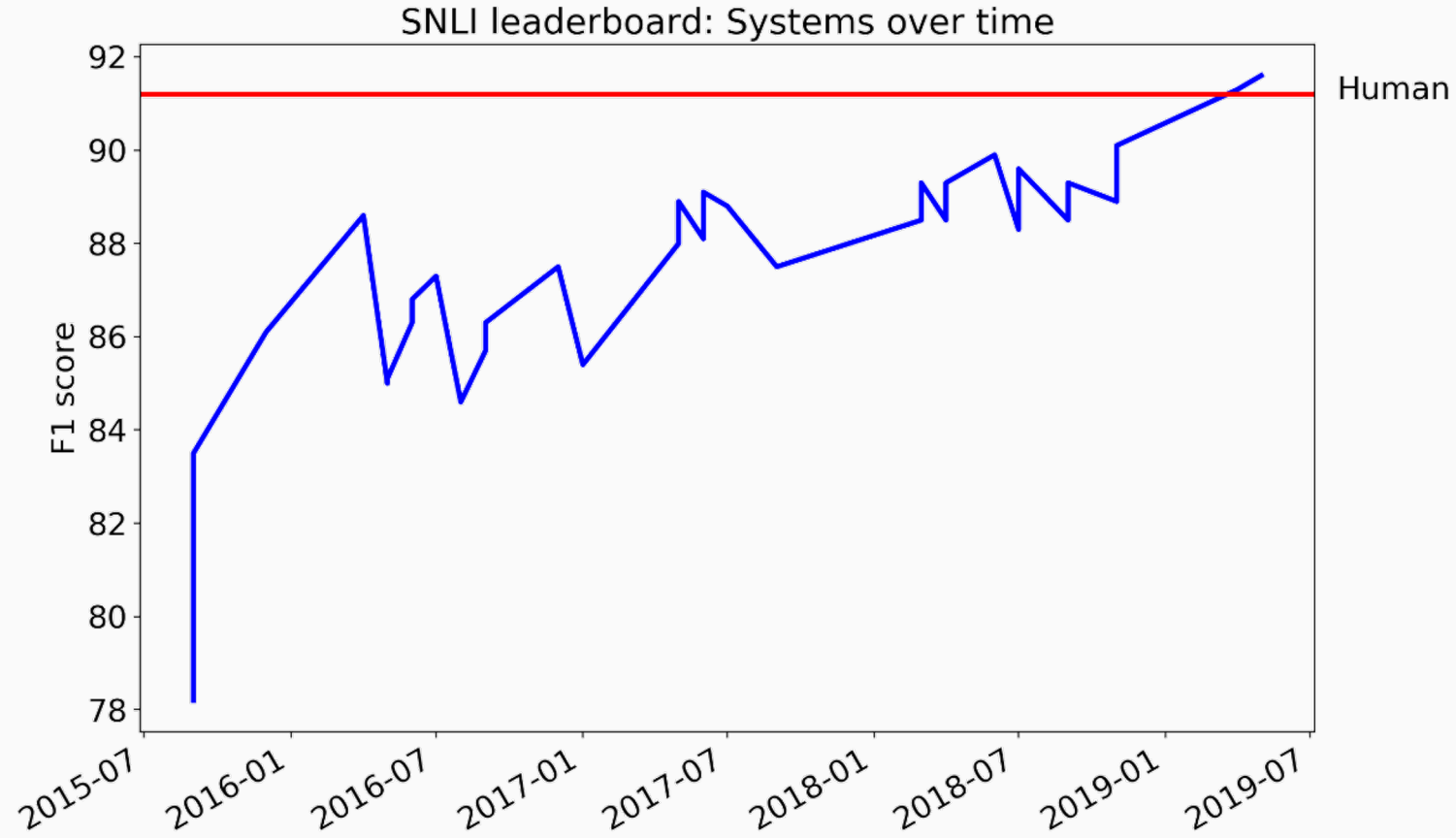
Write a sentence which may be true given the caption, and may not be.

Definitely incorrect Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."*

Write a sentence which contradicts the caption.

Problems (optional) *If something is wrong with the caption that makes it difficult to understand, do your best above and let us know here.*

Models over the SNLI dataset have gotten really good



SNLI examples

A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

The MultiNLI dataset: Multiple genres

Williams et al 2018

Train premises drawn from five genres:

- Fiction: works from 1912–2010 spanning many genres
- Government: reports, letters, speeches, etc., from government websites
- The Slate website
- Telephone: the Switchboard corpus
- Travel: Berlitz travel guides

Additional genres just for dev and test (the mismatched condition):

- The 9/11 report
- Face-to-face: The Charlotte Narrative and Conversation Collection
- Fundraising letters
- Non-fiction from Oxford University Press
- Verbatim: articles about linguistics

The MultiNLI dataset: Multiple genres

Williams et al 2018

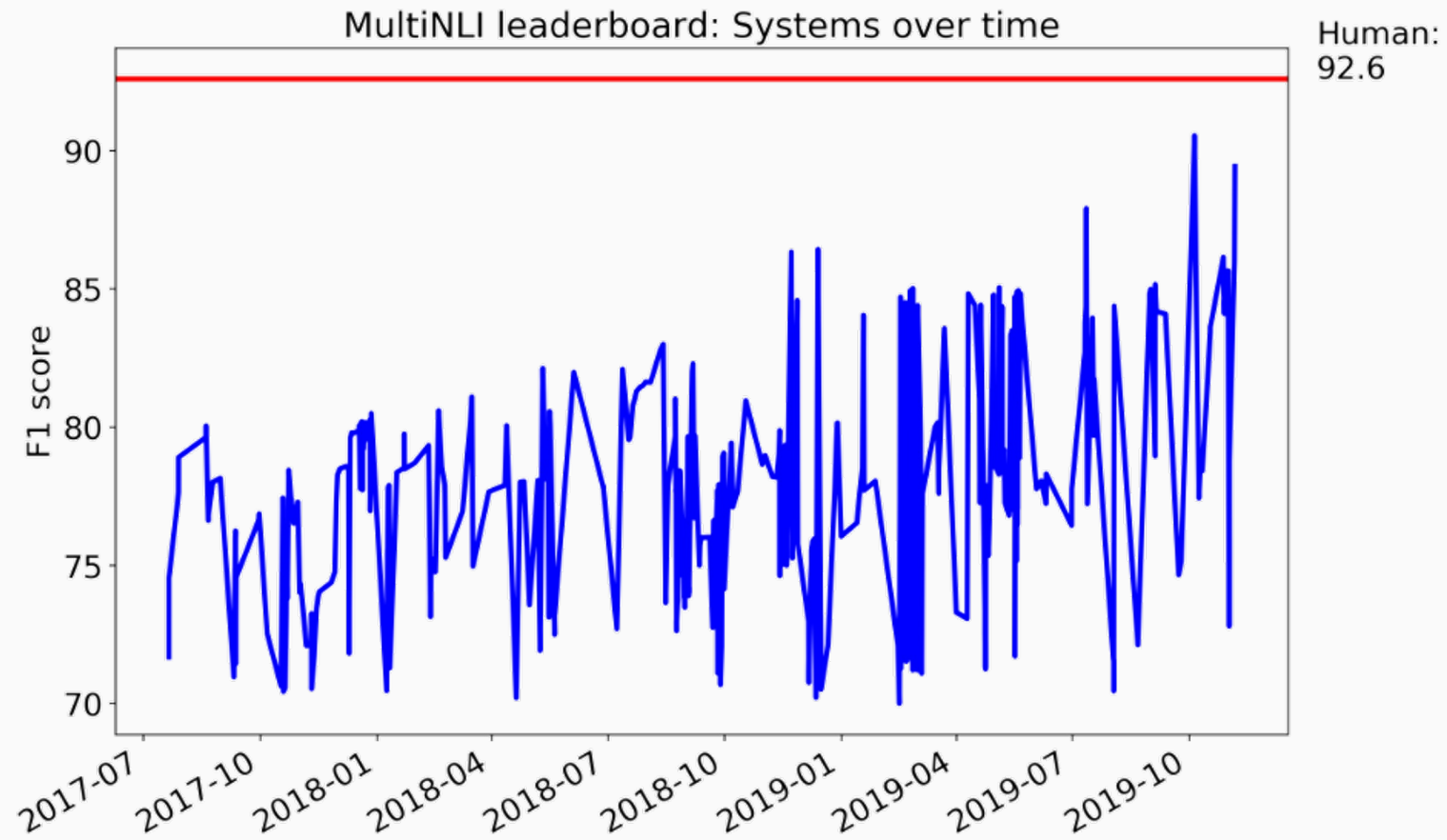
Train premises drawn from five genres:

- Fiction: works from 1912–2010 spanning many genres
- Government: reports, letters, speeches, etc., from government websites
- The Slate website 92,702 train examples; 20K dev; 20K test
- Telephone: the Switchboard corpus 19,647 examples validated by four additional annotators
- Travel: Berlitz travel guides 58.2% examples with unanimous gold label
92.6% of gold labels match the author's label







Additional genres just for dev and test (the mismatched condition):

- The 9/11 report
- Face-to-face: The Charlotte Narrative and Conversation Collection
- Fundraising letters
- Non-fiction from Oxford University Press
- Verbatim: articles about linguistics

MultiNLI progress



NLI tasks are part of the GLUE benchmark

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	Microsoft Alexander v-team	Turing ULR v6		91.3	73.3	97.5	94.2/92.3	93.5/93.1	76.4/90.9	92.5	92.1	96.7	93.6	97.9	55.4
2	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7	92.4	97.9	51.4
3	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9	57.0
4	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7	93.2	96.6	53.3
5	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9	51.7
6	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5	95.2	49.1
7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.2	94.5	53.2
8	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5	52.6
9	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
10	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

Adversarial NLI: Making the dataset more difficult

Nie et al. 2019

- 62,865 labeled examples
- The premises come from diverse sources.
- The hypotheses are written by crowdworkers with the explicit goal of fooling state-of-the-art models.

Adversarial NLI: Making the dataset more difficult

Nie et al. 2019

- 62,865 labeled examples
- The premises come from diverse sources.
- The hypotheses are written by crowdworkers with the explicit goal of fooling state-of-the-art models.
 1. The annotator is presented with a premise sentence and a condition (entailment, contradiction, neutral).
 2. The annotator writes a hypothesis.
 3. A state-of-the-art model makes a prediction about the premise–hypothesis pair.
 4. If the model’s prediction matches the condition, the annotator returns to step 2 to try again.
 5. If the model was fooled, the premise–hypothesis pair is independently validated by other annotators.

Outline

- The task definition
- Datasets and models
- Examples

Natural Language Inference dataset: SNLI

Classify relationship of a premise and hypothesis into 3 classes:
Entailment, *Contradiction*, *Neutral*

Premise Two women are embracing while holding to go packages.

Hypothesis The men are fighting outside a deli.

From the SNLI dataset

Constructed from Image captions (examples talk about scenes)

Natural Language Inference dataset: SNLI

Classify relationship of a premise and hypothesis into 3 classes:
Entailment, *Contradiction*, *Neutral*

Premise Two women are embracing while holding to go packages.

Hypothesis The men are fighting outside a deli.
(*Contradiction*)

From the SNLI dataset

Constructed from Image captions (examples talk about scenes)

Natural Language Inference dataset: MNLI

Classify relationship of a premise and hypothesis into 3 classes:
Entailment, *Contradiction*, *Neutral*

Premise In reviewing this history, it's important to make some crucial distinctions.

Hypothesis Making certain distinctions is imperative in looking back on the past.

From the MNLI dataset

Multi-genre corpus using both spoken and written text.

Natural Language Inference dataset: MNLI

Classify relationship of a premise and hypothesis into 3 classes:
Entailment, *Contradiction*, *Neutral*

Premise In reviewing this history, it's important to make some crucial distinctions.

Hypothesis Making certain distinctions is imperative in looking back on the past. (*Entailment*)

From the MNLI dataset

Multi-genre corpus using both spoken and written text.

Natural Language Inference dataset: Dialogue NLI

Classify relationship of a premise and hypothesis into 3 classes:

Entailment, *Contradiction*, *Neutral*

Premise No politics for me. I would prefer a good heart concert instead.

Hypothesis I have three children all girls.

From the Dialogue NLI dataset

Constructed to test consistency and inference capabilities of dialogue models

Natural Language Inference dataset: Dialogue NLI

Classify relationship of a premise and hypothesis into 3 classes:
Entailment, *Contradiction*, *Neutral*

Premise No politics for me. I would prefer a good heart concert instead.

Hypothesis I have three children all girls.
(Neutral)

From the Dialogue NLI dataset

Constructed to test consistency and inference capabilities of dialogue models

Natural Language Inference dataset: HANS NLI

Classify relationship of a premise and hypothesis into 3 classes:
Entailment, *Contradiction*, *Neutral*

Premise The president advised the doctor.

Hypothesis The doctor advised the president.

From the HANS NLI dataset

To diagnose NLI models for shortcuts (Lexical Overlap above)

Natural Language Inference dataset: HANS NLI

Classify relationship of a premise and hypothesis into 3 classes:
Entailment, *Contradiction*, *Neutral*

Premise The president advised the doctor.

Hypothesis The doctor advised the president.
(Not Entailment)

From the HANS NLI dataset

To diagnose NLI models for shortcuts (Lexical Overlap above)

Natural Language Inference dataset: Breaking NLI

Classify relationship of a premise and hypothesis into 3 classes:

Entailment, *Contradiction*, *Neutral*

Premise The cat sat on the mat.

Hypothesis The cat did not sit on the mat.

From the Breaking NLI dataset

To understand their inference capabilities (negation shown)

Natural Language Inference dataset: Breaking NLI

Classify relationship of a premise and hypothesis into 3 classes:
Entailment, *Contradiction*, *Neutral*

Premise The cat sat on the mat.

Hypothesis The cat did not sit on the mat.
(*Contradiction*)

From the Breaking NLI dataset

To understand their inference capabilities (negation shown)

Summary

- The Textual Entailment/ Natural Language inference task
- Several datasets exist for the task
- The standard approach for building models today
 - Train an encoder model to predict one of the three classes
- What are some problems with the definition of the task? Do we have a full fledged test of reasoning here?