Deep Learning for Natural Language Inference

NAACL-HLT 2019 Tutorial

Follow the slides:
nlitutorial.github.io

Sam Bowman
NYU (New York)

Xiaodan Zhu
Queen's University, Canada

Introduction

Motivations of the Tutorial

Overview

Starting Questions ...

Outline

NLI: What and Why (SB)

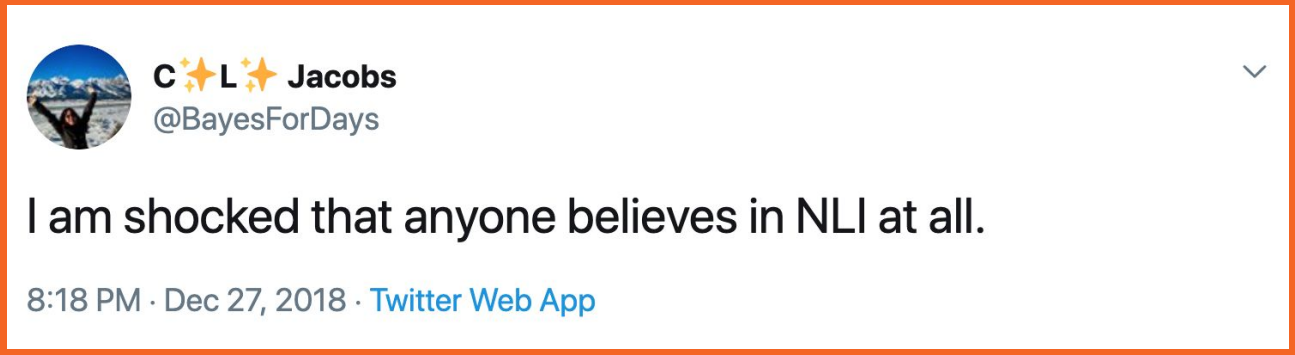
Data for NLI (SB)

Some Methods (SB)

Deep Learning Models (XZ)

- Full Models
- ---(Break, roughly at 10:30)---
- Sentence Vector Models
- Selected Topics

Applications (SB)

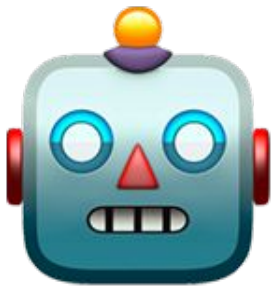


Natural Language Inference: What and Why

Why NLI?

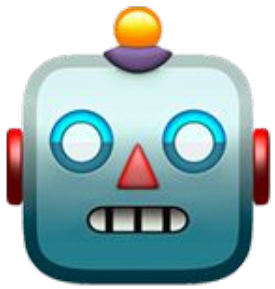
My take, as someone interested in *natural language understanding*...

The Motivating Questions



Can current neural network methods learn to do anything that resembles *compositional semantics*?

The Motivating Questions



Can current neural network methods learn to do anything that resembles *compositional semantics*?

If we take this as a *goal to work toward*, what's our metric?

One possible answer:

Natural Language Inference (NLI)

also known as
recognizing textual entailment (RTE)

"Premise" or "Text" or "Sentence A"

i'm not sure what the overnight low was

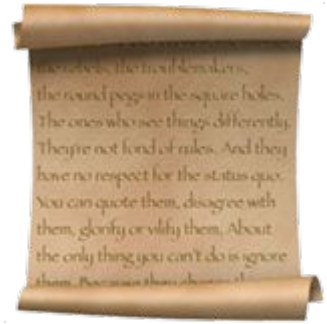
{**entails**, contradicts, neither}

"Hypothesis" or "Sentence B"

I don't know how cold it got last night.

[Dagan et al. '05](), [MacCartney '09]()
Example from [MNLI]()

A Definition



We say that T entails H if, typically, a human reading T would infer that H is most likely true.

- Ido Dagan '05

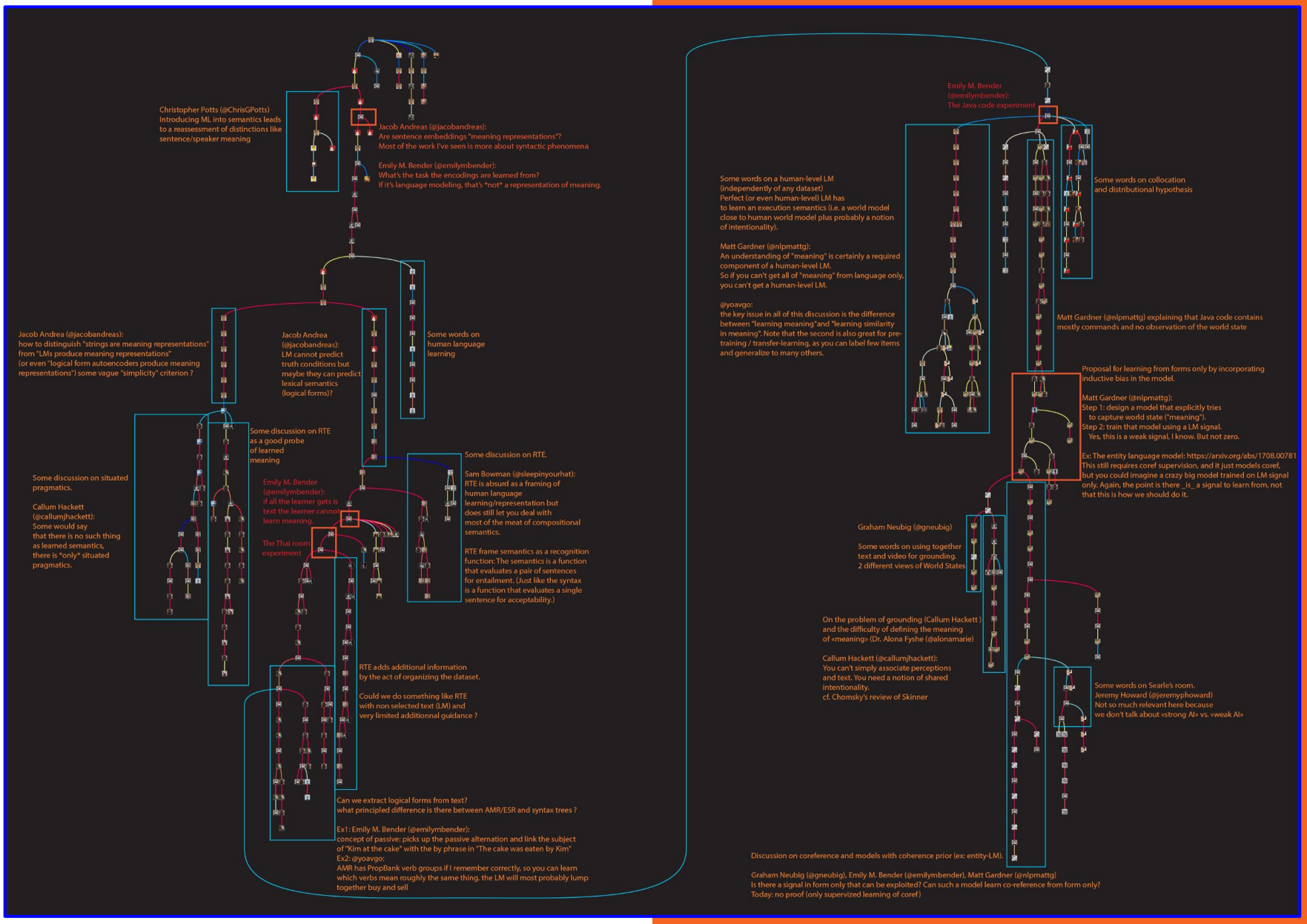
(See Manning '06 for discussion.)

The Big Question

What kind of a thing is the meaning of a sentence?

The Big Question

~~What kind of a thing is the meaning of a sentence?~~


Christopher Potts (@ChrisGPotts)
Introducing ML into semantics leads to a reassessment of distinctions like sentence/speaker meaning
Jacob Andreas (@jacobandreas):
Are sentence embeddings "meaning representations"?
Most of the work I've seen is more about syntactic phenomena
Emily M. Bender (@emilymbender):
What's the task the encodings are learned from?
If it's language modeling, that's *not* a representation of meaning.
Jacob Andrea (@jacobandreas):
how to distinguish "strings are meaning representations" from "LMs produce meaning representations" (or even "logical form autoencoders produce meaning representations") some vague "simplicity" criterion ?
Jacob Andrea (@jacobandreas):
LM cannot predict truth conditions but maybe they can predict lexical semantics (logical forms)?
Some words on human language learning
Some discussion on RTE as a good probe of learned meaning
Some discussion on situated pragmatics.
Callum Hackett (@callumjhackett):
Some would say that there is no such thing as learned semantics, there is *only* situated pragmatics.
Emily M. Bender (@emilymbender):
if all the learner gets is text the learner cannot learn meaning.
The Thai room experiment
Some discussion on RTE.
Sam Bowman (@sleepinyourhat):
RTE is absurd as a framing of human language learning/representation but does still let you deal with most of the meat of compositional semantics.
RTE frame semantics as a recognition function: The semantics is a function that evaluates a pair of sentences for entailment. (Just like the syntax is a function that evaluates a single sentence for acceptability.)
RTE adds additional information by the act of organizing the dataset.
Could we do something like RTE with non selected text (LM) and very limited additionnal guidance ?
Can we extract logical forms from text?
what principled difference is there between AMR/ESR and syntax trees ?
Ex1: Emily M. Bender (@emilymbender):
concept of passive: picks up the passive alternation and link the subject of "Kim at the cake" with the by phrase in "The cake was eaten by Kim"
Ex2: @yoavgo:
AMR has PropBank verb groups if I remember correctly, so you can learn which verbs mean roughly the same thing, the LM will most probably lump together buy and sell
Emily M. Bender (@emilymbender):
The Java code experiment
Some words on a human-level LM (independently of any dataset)
Perfect (or even human-level) LM has to learn an execution semantics (i.e. a world model close to human world model plus probably a notion of intentionality).
Matt Gardner (@nlpmattg):
An understanding of "meaning" is certainly a required component of a human-level LM.
So if you can't get all of "meaning" from language only, you can't get a human-level LM.
@yoavgo:
the key issue in all of this discussion is the difference between "learning meaning"and "learning similarity in meaning". Note that the second is also great for pre-training / transfer-learning, as you can label few items and generalize to many others.
Some words on collocation and distributional hypothesis
Matt Gardner (@nlpmattg) explaining that Java code contains mostly commands and no observation of the world state
Proposal for learning from forms only by incorporating inductive bias in the model.
Matt Gardner (@nlpmattg):
Step 1: design a model that explicitly tries to capture world state ("meaning").
Step 2: train that model using a LM signal.
Yes, this is a weak signal, I know. But not zero.
Ex: The entity language model: https://arxiv.org/abs/1708.00781
This still requires coref supervision, and it just models coref, but you could imagine a crazy big model trained on LM signal only. Again, the point is there _is_ a signal to learn from, not that this is how we should do it.
Graham Neubig (@gneubig)
Some words on using together text and video for grounding.
2 different views of World States
On the problem of grounding (Callum Hackett) and the difficulty of defining the meaning of «meaning» (Dr. Alona Fyshe (@alonamarie)
Callum Hackett (@callumjhackett):
You can't simply associate perceptions and text. You need a notion of shared intentionality.
cf. Chomsky's review of Skinner
Some words on Searle's room.
Jeremy Howard (@jeremyphoward)
Not so much relevant here because we don't talk about «strong AI» vs. «weak AI»
Discussion on coreference and models with coherence prior (ex: entity-LM).
Graham Neubig (@gneubig), Emily M. Bender (@emilymbender), Matt Gardner (@nlpmattg)
Is there a signal in form only that can be exploited? Can such a model learn co-reference from form only?
Today: no proof (only supervized learning of coref)

The Big Question

~~What kind of a thing is the meaning of a sentence?~~

The Big Question

~~What kind of a thing is the meaning of a sentence?~~

What concrete phenomena do you have to deal with to understand a sentence?

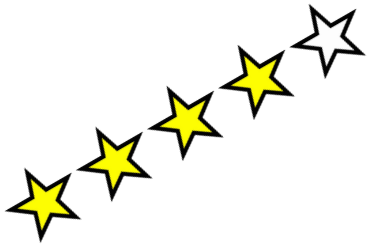
Judging Understanding with NLI

To reliably perform well at NLI, your method for sentence understanding must be able to interpret and use the full range of phenomena we talk about in compositional semantics:*

- Lexical entailment (*cat* vs. *animal*, *cat* vs. *dog*)
- Quantification (*all*, *most*, *fewer than eight*)
- Lexical ambiguity and scope ambiguity (*bank*, ...)
- Modality (*might*, *should*, ...)
- Common sense background knowledge

…

* without grounding to the outside world.



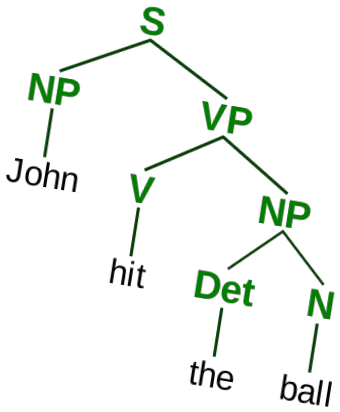
Why not Other Tasks?

Many tasks that have been used to evaluate sentence representation models don't require models to deal with the full complexity of compositional semantics:

- Sentiment analysis
- Sentence similarity

…

Why not Other Tasks?



NLI is one of many NLP tasks that require robust compositional sentence understanding:

- Machine translation
- Question answering
- Goal-driven dialog
- Semantic parsing
- Syntactic parsing
- Image–caption matching

…

But it's the simplest of these.

Detour:
Entailments and Truth Conditions

?

Most formal semantics research (and some semantic parsing research) deals with truth conditions.

See Katz ‘72

Detour: **Entailments and Truth Conditions**

?

Most formal semantics research (and some semantic parsing research) deals with truth conditions.

In this view understanding a sentence means (roughly) **characterizing the set of situations in which that sentence is true.**

See Katz '72

Detour: **Entailments and Truth Conditions**

?

Most formal semantics research (and some semantic parsing research) deals with truth conditions.

In this view understanding a sentence means (roughly) **characterizing the set of situations in which that sentence is true.**

This requires some form of grounding:

Truth-conditional semantics is strictly harder than NLI.

See Katz '72

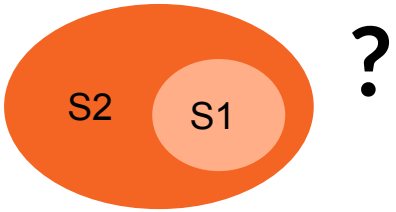
Detour:
Entailments and Truth Conditions

?

If you know the truth conditions of two sentences, can you work out whether one entails the other?

See Katz ‘72

Detour:
Entailments and Truth Conditions



?

If you know the truth conditions of two sentences, can you work out whether one entails the other?

See Katz '72

Detour:
Entailments and Truth Conditions

?

Can you work out whether one sentence entails another without knowing their truth conditions?

See Katz ‘72

Detour:
Entailments and Truth Conditions

Isobutylphenylpropionic acid is a medicine for headaches.

{entails, contradicts, neither}**?**

Isobutylphenylpropionic acid is a medicine.

Can you work out whether one sentence entails another without knowing their truth conditions?

See Katz '72

Another set of motivations...

Question Answering: Given a question (premise), identify a text that entails an answer (hypothesis).

Information Retrieval: Given a query (hypothesis), identify texts that entail that query (premises).

Summarization: Given a text (premise) *T*, create or identify a text that *T* entails.

Summarization: Omit sentences that are entailed by others.

Machine translation: Mutual entailment between texts in different languages.

We'll revisit this later!

-Bill MacCartney, Stanford CS224U Slides

Natural Language Inference: **Data**

...an incomplete survey

FraCaS Test Suite

P: *No delegate finished the report.*
H: *Some delegate finished the report on time.*
Label: *no entailment*

- 346 examples
- Manually constructed by experts
- Target strict logical entailment

Cooper et al. ‘96, MacCartney ‘09

Recognizing Textual Entailment (RTE) 1–7

P: *Cavern Club sessions paid the Beatles £15 evenings and £5 lunchtime.*
H: *The Beatles perform at Cavern Club at lunchtime.*
Label: entailment

- Seven annual competitions (First PASCAL, then NIST)
- Some variation in format, but about 5000 NLI-format examples total
- Premises (*texts*) drawn from naturally occurring text, often long/complex
- Expert-constructed hypotheses

Dagan et al. ‘06 et seq.

Sentences Involving Compositional Knowledge (SICK)

P: *The brown horse is near a red barrel at the rodeo*
H: *The brown horse is far from a red barrel at the rodeo*
Label: contradiction

- Corpus for a 2014 SemEval shared task competition
- Deliberately restricted task: No named entities, idioms, etc.
- Pairs created by semi-automatic manipulation rules on image and video captions
- About 10,000 examples, labeled for entailment *and* semantic similarity (1–5 scale)

Marelli et al. ‘14

The Stanford NLI Corpus (SNLI)

P: *A black race car starts up in front of a crowd of people.*
H: *A man is driving down a lonely road.*
Label: contradiction

- Premises derived from image captions (Flickr 30k), hypotheses created by crowdworkers
- About 550,000 examples; first NLI corpus to see encouraging results with neural networks

Bowman et al. ‘15

Multi-Genre NLI (MNLI)

P: *yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual*
H: *August is a black out month for vacations in the company.*
Label: contradiction

- Multi-genre follow-up to SNLI: Premises come from ten different sources of written and spoken language (mostly via OpenANC), hypotheses written by crowdworkers
- About 400,000 examples

Williams et al. ‘18

Multi-Premise Entailment (MPE)

Premises:

1. Three men are working construction on top of a building.
2. Three male construction workers on a roof working in the sun.
3. One man is shirtless while the other two men work on construction.
4. Two construction workers working on infrastructure, while one worker takes a break.

Hypothesis:

A man smoking a cigarette. ⇒NEUTRAL

- *Multi-premise entailment* from a set of sentences describing a scene
- Derived from Flickr30k image captions
- About 10,000 examples

Lai et al. ‘17

Crosslingual NLI (XNLI)

P: 让我告诉你，美国人最终如何看待你作为独立顾问的表现。
H: 美国人完全不知道您是独立律师。
Label: contradiction

- A new *development and test* set for MNLI, translated into 15 languages
- About 7,500 examples per language
- Meant to evaluate cross-lingual transfer: Train on English MNLI, evaluate on another target language(s)
- Sentences translated one-by-one, so some inconsistencies

Conneau et al. ‘18

Crosslingual NLI (XNLI)

P: 让我告诉你，美国人最终如何看待你作为独立顾问的表现。
H: 美国人完全不知道您是独立律师。
Label: contradiction

- A new *development and test* set for MNLI, translated into 15 languages
- About 7,500 examples per language
- Meant to evaluate cross-lingual transfer: Train on English MNLI, evaluate on another target language(s)
- Sentences translated one-by-one, so some inconsistencies

Conneau et al. ‘18

SciTail

P: *Cut plant stems and insert stem into tubing while stem is submerged in a pan of water.*
H: Stems transport water to other parts of the plant through a system of tubes.
Label: neutral

- Created by pairing statements from science tests with information from the web
- First NLI set built entirely on *existing* text
- About 27,000 pairs

Khot et al. ‘18

In Depth:
SNLI and MNLI

First:
Entity and Event Coreference in NLI

One event or two?

Premise: *A boat sank in the Pacific Ocean.*

Hypothesis: *A boat sank in the Atlantic Ocean.*

One event or two? One.

Premise: *A boat sank in the Pacific Ocean.*

Hypothesis: *A boat sank in the Atlantic Ocean.*

Label: contradiction

One event or two?

Premise: *Ruth Bader Ginsburg was appointed to the US Supreme Court.*

Hypothesis: *I had a sandwich for lunch today*

One event or two? Two.

Premise: *Ruth Bader Ginsburg was appointed to the US Supreme Court.*

Hypothesis: *I had a sandwich for lunch today*

Label: neutral

But if we allow for this, then can we ever get a contradiction between two natural sentences?

One event or two? Two.

Premise: *A boat sank in the Pacific Ocean.*

Hypothesis: *A boat sank in the Atlantic Ocean.*

Label: neutral

One event or two? One, always.

Premise: *A boat sank in the Pacific Ocean.*

Hypothesis: *A boat sank in the Atlantic Ocean.*

Label: contradiction

How do we turn tricky constraint this into something annotators can learn quickly?

One event or two? One, always.

Premise: *Ruth Bader Ginsburg was appointed to the US Supreme Court.*

Hypothesis: *I had a sandwich for lunch today*

Label: *contradiction*

One *photo* or two? One, always.

Premise: Ruth Bader Ginsburg being appointed to the US Supreme Court.

Hypothesis: A man eating a sandwich for lunch.

Label: can't be the same photo (so: contradiction)

Our Solution:
The SNLI Data Collection Prompt

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** a **false** description of the photo.

Photo caption **An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.**

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."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.

Problems (optional) *If something is wrong, have a look at the FAQ, do your best above, and let us know here.*

Source captions from Flickr30k: Young, et al. '14

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** a **false** description of the photo.

Photo caption An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.

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. **Entailment**

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."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.

Problems (optional) *If something is wrong, have a look at the FAQ, do your best above, and let us know here.*

Source captions from Flickr30k: Young, et al. '14

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** a **false** description of the photo.

Photo caption **An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.**

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. **Entailment**

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. **Neutral**

Definitely incorrect Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.

Problems (optional) *If something is wrong, have a look at the FAQ, do your best above, and let us know here.*

Source captions from Flickr30k: Young, et al. '14

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** a **false** description of the photo.

Photo caption An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.

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. **Entailment**

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. **Neutral**

Definitely incorrect Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption. **Contradiction**

Problems (optional) *If something is wrong, have a look at the FAQ, do your best above, and let us know here.*

Source captions from Flickr30k: Young, et al. '14

What we got

Some sample results

Premise: *Two women are embracing while holding to go packages.*

Hypothesis: *Two woman are holding packages.*

Label: Entailment

Some sample results

Premise: *A man in a blue shirt standing in front of a garage-like structure painted with geometric designs.*

Hypothesis: *A man is repainting a garage*

Label: Neutral

MNLI

MNLI

- Same intended definitions for labels: Assume coreference.
- More genres—not just concrete visual scenes.
- Needed more complex annotator guidelines and more careful quality control, but reached same level of annotator agreement.

What we got

Typical Dev Set Examples

Premise: *In contrast, suppliers that have continued to innovate and expand their use of the four practices, as well as other activities described in previous chapters, keep outperforming the industry as a whole.*

Hypothesis: *The suppliers that continued to innovate in their use of the four practices consistently underperformed in the industry.*

Label: Contradiction

Genre: Oxford University Press (Nonfiction books)

Typical Dev Set Examples

Premise: *someone else noticed it and i said well i guess that's true and it was somewhat melodious in other words it wasn't just you know it was really funny*

Hypothesis: *No one noticed and it wasn't funny at all.*

Label: Contradiction

Genre: Switchboard (Telephone Speech)

Key Figures

Tag	SNLI	MultiNLI
Pronouns (PTB)	34	**68**
Quantifiers	33	**63**
Modals (PTB)	<1	**28**
Negation (PTB)	5	**31**
'Wh' Words (PTB)	5	**30**
Belief Verbs	<1	**19**
Time Terms	19	**36**
Conversational Pivots	<1	**14**
Presupposition Triggers	8	**22**
Comparatives/Superlatives (PTB)	3	**17**
Conditionals	4	**15**
Tense Match (PTB)	62	**69**
Interjections (PTB)	<1	**5**
>20 Words	<1	**5**
Existentials (PTB)	5	**8**

The Train-Test Split

The MNLI Corpus

Genre	Train	Dev	Test
Captions (SNLI Corpus)	(550,152)	(10,000)	(10,000)
Fiction	77,348	2,000	2,000
Government	77,350	2,000	2,000
Slate	77,306	2,000	2,000
Switchboard (Telephone Speech)	83,348	2,000	2,000
Travel Guides	77,350	2,000	2,000

The MNLI Corpus

Genre	Train	Dev	Test
Captions (SNLI Corpus)	(550,152)	(10,000)	(10,000)
Fiction	77,348	2,000	2,000
Government	77,350	2,000	2,000
Slate	77,306	2,000	2,000
Switchboard (Telephone Speech)	83,348	2,000	2,000
Travel Guides	77,350	2,000	2,000
9/11 Report	0	2,000	2,000
Face-to-Face Speech	0	2,000	2,000
Letters	0	2,000	2,000
OUP (Nonfiction Books)	0	2,000	2,000
Verbatim (Magazine)	0	2,000	2,000
Total	**392,702**	**20,000**	**20,000**

The MNLI Corpus

Genre	Train	Dev	Test	
Captions (SNLI Corpus)	(550,152)	(10,000)	(10,000)	
Fiction	77,348	2,000	2,000	
Government	77,350	2,000	2,000	
Slate	77,306	2,000	2,000	*genre-matched evaluation*
Switchboard (Telephone Speech)	83,348	2,000	2,000	
Travel Guides	77,350	2,000	2,000	
9/11 Report	0	2,000	2,000	
Face-to-Face Speech	0	2,000	2,000	
Letters	0	2,000	2,000	*genre-mismatched evaluation*
OUP (No			00	
Verba				
Total				

Good news:

Most models perform similarly on both sets!

Annotation Artifacts

Annotation Artifacts

For SNLI:

P: ???

H: *Someone is not crossing the road.*

Label: entailment, contradiction, neutral?

Poliak et al. ‘18, Tsuchiya ‘18, Gururangan et al. ‘18

Annotation Artifacts

For SNLI:

P: ???

H: *Someone is not crossing the road.*

Label: entailment, **contradiction**, neutral?

Poliak et al. ‘18, Tsuchiya ‘18, Gururangan et al. ‘18

Annotation Artifacts

For SNLI:

P: ???

H: *Someone is not crossing the road.*

Label: entailment, **contradiction**, neutral?

P: ???

H: *Someone is outside.*

Label: entailment, contradiction, neutral?

Poliak et al. ‘18, Tsuchiya ‘18, Gururangan et al. ‘18

Annotation Artifacts

For SNLI:

P: ???

H: *Someone is not crossing the road.*

Label: entailment, **contradiction**, neutral?

P: ???

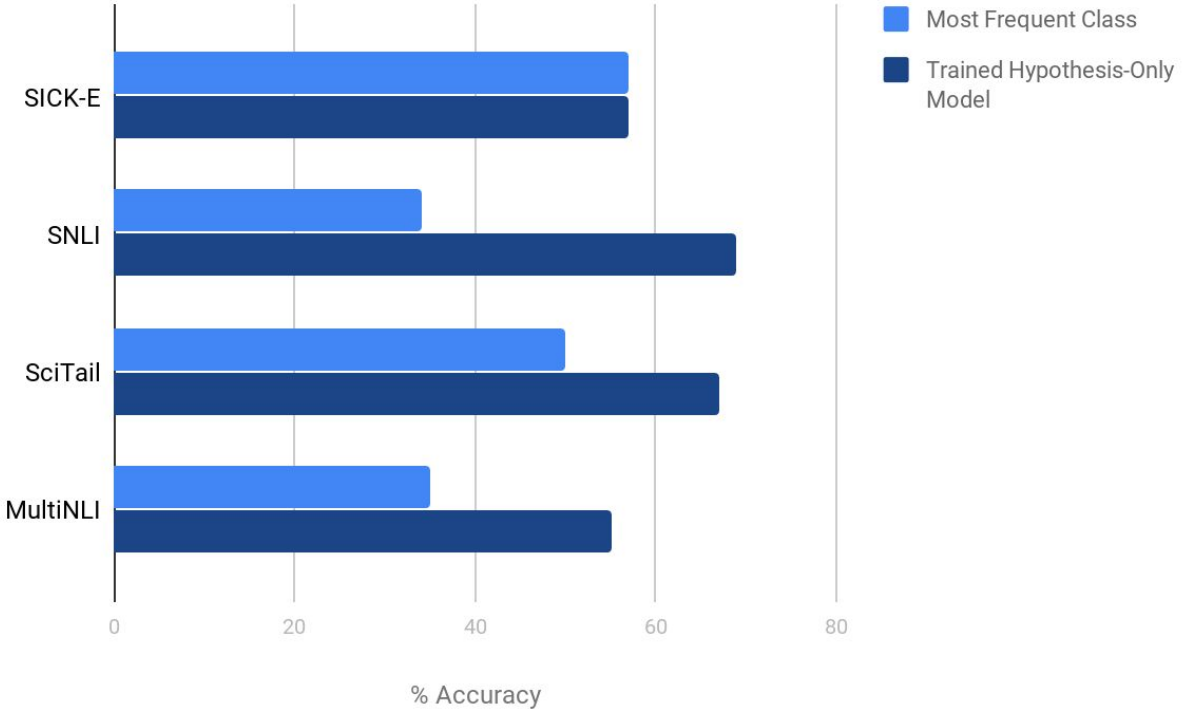
H: *Someone is outside.*

Label: entailment, contradiction, neutral?

Poliak et al. ‘18, Tsuchiya ‘18, Gururangan et al. ‘18

Annotation Artifacts

Models can do moderately well on NLI datasets without looking at the hypothesis!

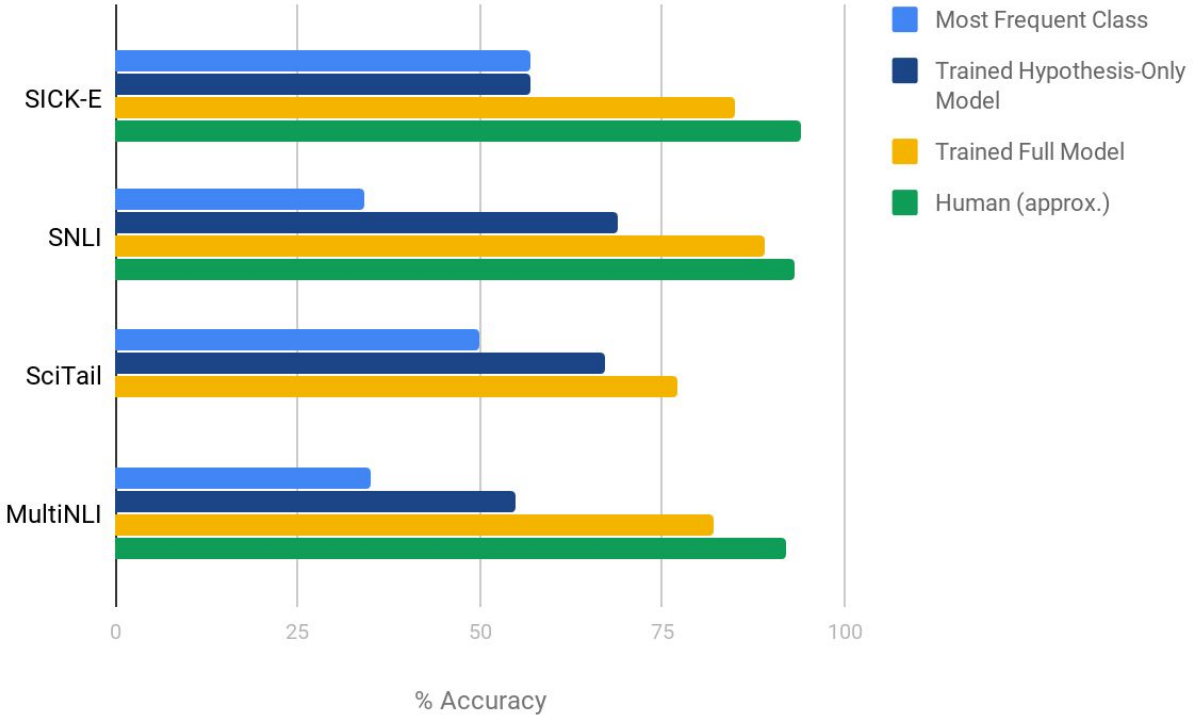


Single-genre SNLI especially vulnerable. SciTail not immune.

Poliak et al. ‘18 (source of numbers), Tsuchiya ‘18, Gururangan et al. ‘18

Annotation Artifacts

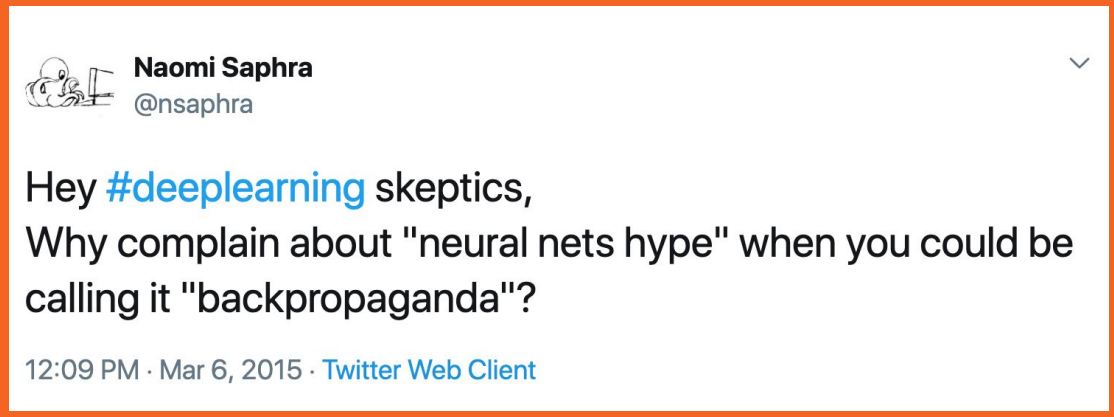
Models can do moderately well on NLI datasets without looking at the hypothesis!



...but hypothesis-only models are still far below ceiling.

These datasets are easier than they look, but not trivial.

Poliak et al. ‘18 (source of numbers), Tsuchiya ‘18, Gururangan et al. ‘18



Natural Language Inference: **Some Methods**

(This is not the deep learning part.)

Feature-Based Models

Some earlier NLI work involved learning with shallow features:

- Bag of words features on hypothesis
- Bag of word-pairs features to capture alignment
- Tree kernels
- Overlap measures like BLEU

These methods work surprisingly well, but not competitive on current benchmarks.

MacCartney ‘09, Stern and Dagan ‘12, Bowman et al. ‘15

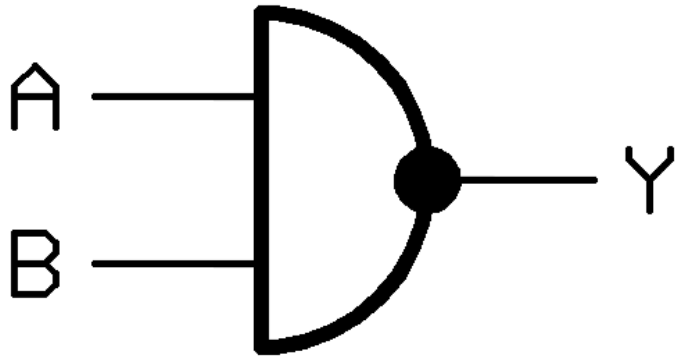
Natural Logic

Much non-ML work on NLI involves natural logic:

- A formal logic for deriving entailments between sentences.
- Operates directly on parsed sentences (*natural* language), no explicit logical forms.
- Generally sound but far from complete—only supports inferences between sentences with clear structural parallels.
- Most NLI datasets aren't strict logical entailment, and require some unstated premises—this is hard.

Lakoff '70, Sánchez Valencia '91, MacCartney '09, Icard III & Moss '14, Hu et al. '19

Theorem Proving



Another thread of work has attempted to translate sentences into *logical forms* (semantic parsing) and use theorem proving methods to find valid inferences.

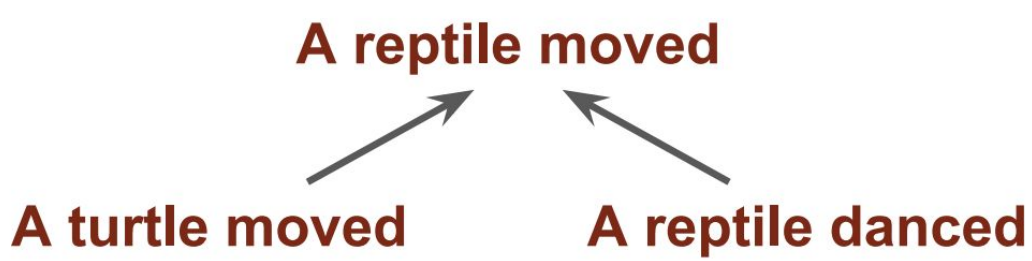
- Open-domain semantic parsing is still hard!
- Unstated premises and common sense can still be a problem.

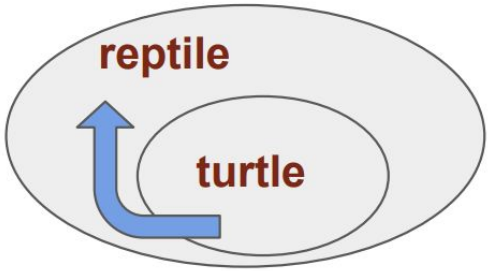
Bos and Markert ‘05, Beltagy et al. ‘13, Abzianidze ‘17

In Depth:
Natural Logic

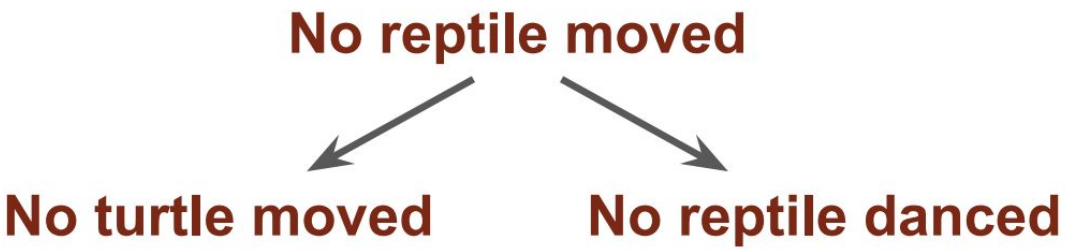
Monotonicity

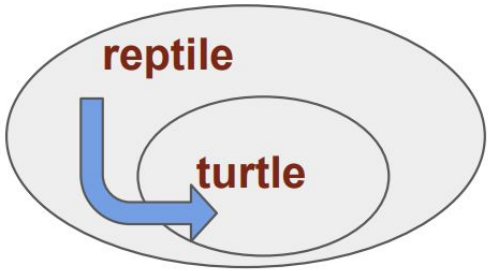
Upward monotone: preserve entailments from **subsets** to **supersets**:





Downward monotone: preserve entailments from **supersets** to **subsets**:





Non-monotone: do not preserve entailment in either direction.

Upward monotonicity in language

- Upward monotonicity is sort of the default for lexical items
- Most determiners (e.g., *a, some, at least, more than*)
- The **second** argument of *every* (*danced* in *every turtle danced*)
- Positive implicatives (e.g., *manage to, succeed to, force to*)

Downward monotonicity in language

- Negations (e.g., *not, n't, never, no, nothing, nowhere, none, neither*)
- The **first** argument of *every* (*turtle* in *every turtle danced*)
- Determiners like *at most, few, fewer/less than*
- Conditional antecedents (*if*-clauses)
- Negative implicatives (e.g., *forget to, refuse to, hesitate to*)
- Negative attitude verbs like *doubt* and *deny* (at least approximately)
- Adverbs like *rarely* and *hardly*

Monotonicity features

- Edits that broaden/weaken preserve forward entailment:
  - Deleting modifiers
  - Changing specific terms to more general ones.
  - Dropping conjuncts, adding disjuncts.
- Edits that narrow/strengthen do not preserve forward entailment:
  - Adding modifiers
  - Changing general terms to specific ones.
  - Adding conjuncts, dropping disjuncts.
- In downward monotone environments, the above are **reversed**.

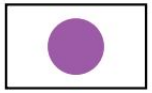
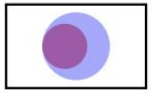
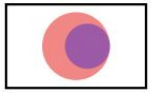
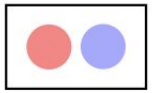
Poll:
Monotonicity

Which of these contexts are upward monotone?

Example: *Some **dogs** are cute*
This is upward monotone, since you can replace *dogs* with a more general term like *animals*, and the sentence must still be true.

1. *Most **cats** meow.*
2. *Some parrots **talk**.*
3. *More than six students **wear purple hats**.*

MacCartney's *Natural Logic* Label Set

	X ≡ Y	equivalence	couch ≡ sofa
	X ⊏ Y	forward entailment	crow ⊏ bird
	X ⊐ Y	reverse entailment	European ⊐ French
	X ^ Y	negation	human ^ non-human
	X | Y	alternation	cat | dog
	X ‿ Y	cover	animal ‿ non-human
	X # Y	independence	hungry # hippo

MacCartney and Manning '09

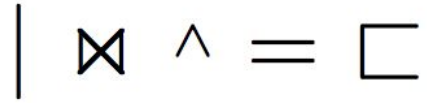
Beyond Up and Down: Projectivity

X | Y	not-X # not-Y	X ⊏ not-Y	not-X ⊐ Y
X = Y	not-X = not-Y	X | not-Y	not-X | Y
X # Y	not-X # not-Y	X # not-Y	not-X # Y
X ⊏ Y	not-X ⊐ not-Y	X | not-Y	not-X # Y
X ⊐ Y	not-X ⊏ not-Y	X # not-Y	not-X | Y

MacCartney and Manning ‘09

Chains of Relations

If we know A | B and B ^ C, what do we know?



So A ⊏ C

⋈	≡	⊏	⊐	^	|	⌣	#
≡	≡	⊏	⊐	^	|	⌣	#
⊏	⊏	⊏	≡⊏⊐|#	|	|	⊏^|⌣#	⊏|#
⊐	⊐	≡⊏⊐⌣#	⊐	⌣	⊐^|⌣#	⌣	⊐⌣#
^	^	⌣	|	≡	⊐	⊏	#
|	|	⊏^|⌣#	|	⊏	≡⊏⊐|#	⊏	⊏|#
⌣	⌣	⌣	⊐^|⌣#	⊐	⊐	≡⊏⊐⌣#	⊐⌣#
#	#	⊏⌣#	⊐|#	#	⊐|#	⊏⌣#	•

MacCartney and Manning '09

Putting it all together

What's the relation between this sentence and the previous sentence?
Use projectivity/monotonicity.

What's the relation between the things we substituted?
Look this up.

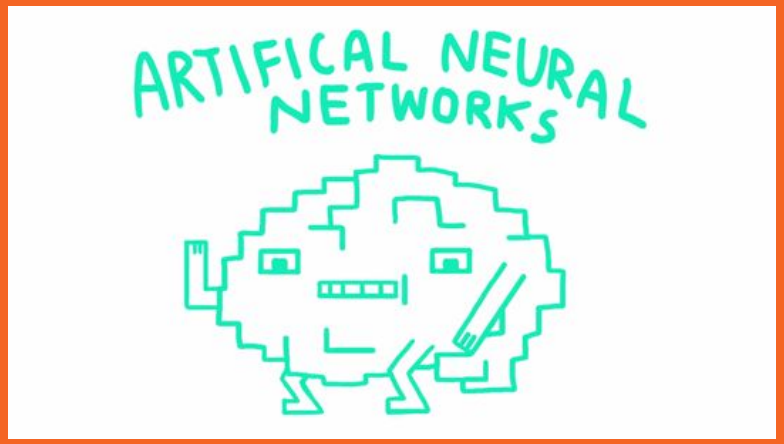
What's the relation between this sentence and the original sentence?
Use join.

i	x_i	e_i	$\beta(e_i)$	$\beta(x_{i-1}, e_i)$	$\beta(x_0, x_i)$
0	*Stimpy is a cat*				
1	*Stimpy is a dog*	SUB(*cat*, *dog*)	|	|	|
2	*Stimpy is not a dog*	INS(*not*)	^	^	⊏
3	*Stimpy is not a poodle*	SUB(*dog*, *poodle*)	⊐	⊏	⊏

MacCartney and Manning '09

Natural Logic: Limitations

- Efficient, *sound* inference procedure, but…
  - ...not *complete*.
- De Morgan's laws for quantifiers:
  - All dogs bark.
  - No dogs don't bark.
- (Plus common sense and unstated premises.)



Natural Language Inference: **Deep Learning Methods**

Deep-Learning Models for NLI

Before we delve into DL models ...

Right, there are many really good reasons we should be excited in DL-based models.

Deep-Learning Models for NLI

Before we delve into DL models ...

Right, there are many really good reasons we should be excited in DL-based models.

But, there are also many good reasons we may want to know nice non-DL research performed before.

Deep-Learning Models for NLI

Before we delve into DL models ...

Right, there are many really good reasons we should be excited in DL-based models.

But, there are also many good reasons we may want to know nice non-DL research performed before.

Also, it is alway intriguing to think how the final NLI models would look like.

Two Categories of Deep Learning Models for NLI

- We roughly organize our discussion on deep learning models for NLI by two typical categories:
  - **Category I**: NLI models that explore both sentence representation and cross-sentence statistics (e.g., attention) to detect NLI relations. (*Full models*)
  - **Category II**: NLI models that do not use cross-sentence attention. (*Sentence-vector-based models*)
    - This category of modelling is of interest because NLI is a good test bed for learning representation for sentences, as discussed earlier in the tutorial.

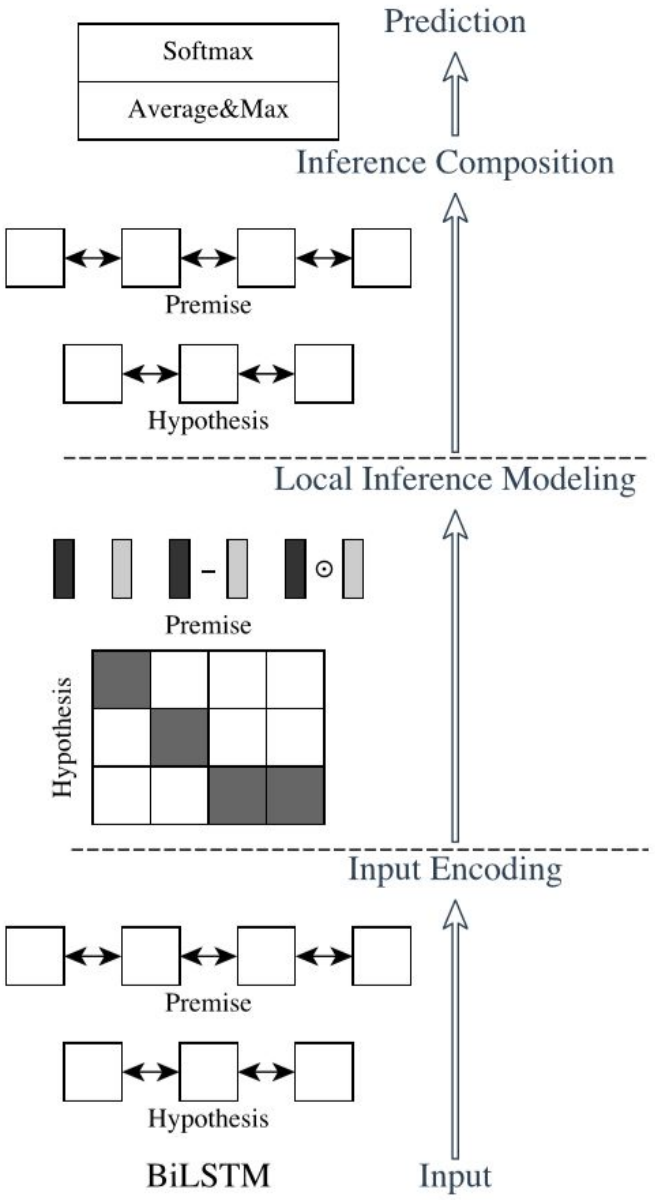
Outline

- "Full" deep-learning models for NLI
  - Baseline models and typical components
  - NLI models enhanced with syntactic structures
  - NLI models considering semantic roles
  - Incorporating external knowledge
    - Incorporating human-curated structured knowledge
    - Leveraging unstructured data with self-supervision (aka. unsupervised pretraining)
- Sentence-vector-based NLI models
  - A top-ranked model in RepEval-2017
  - Current top models based on dynamic self-attention
- Several additional topics

Outline

- **"Full" deep-learning models for NLI**
  - **Baseline models and typical components**
  - NLI models enhanced with syntactic structures
  - NLI models considering semantic roles
  - Incorporating external knowledge
    - Incorporating human-curated structured knowledge
    - Leveraging unstructured data with self-supervision (aka. unsupervised pretraining)
- Sentence-vector-based NLI models
  - A top-ranked model in RepEval-2017
  - Current top models based on dynamic self-attention
- Several additional topics

Enhanced Sequential Inference Models (ESIM)



Layer 3: Inference Composition/Aggregation

Perform composition/aggregation over local inference output to make the global judgement.

Layer 2: Local Inference Modelling

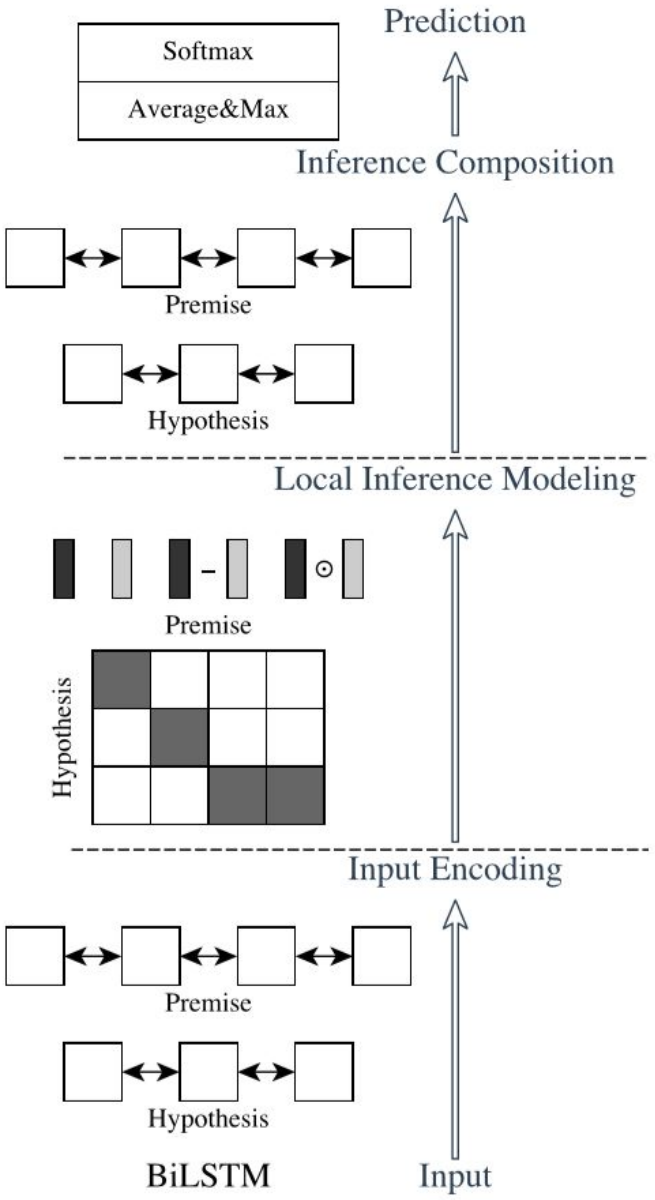
Collect information to perform "local" inference between words or phrases. (Some heuristics works well in this layer.)

Layer 1: Input Encoding

ESIM uses BiLSTM, but different architectures can be used here, e.g., transformer-based, ELMo, densely connected CNN, tree-based models, etc.

Chen et al. '17

Enhanced Sequential Inference Models (ESIM)



Layer 3: Inference Composition/Aggregation

Perform composition/aggregation over local inference output to make the global judgement.

Layer 2: Local Inference Modelling

Collect information to perform "local" inference between words or phrases. (Some heuristics works well in this layer.)

Layer 1: Input Encoding

ESIM uses BiLSTM, but different architectures can be used here, e.g., transformer-based, ELMo, densely connected CNN, tree-based models, etc.

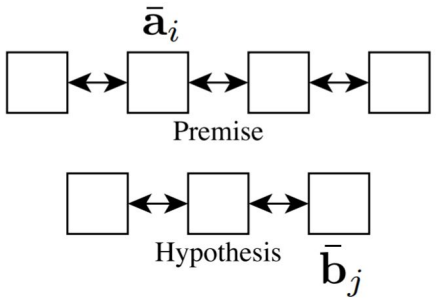
Chen et al. '17

Encoding Premise and Hypothesis

- For a premise sentence, **a**, and a hypothesis sentence **b**:

$$\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_{\ell_a})$$
$$\mathbf{b} = (\mathbf{b}_1, \ldots, \mathbf{b}_{\ell_b})$$

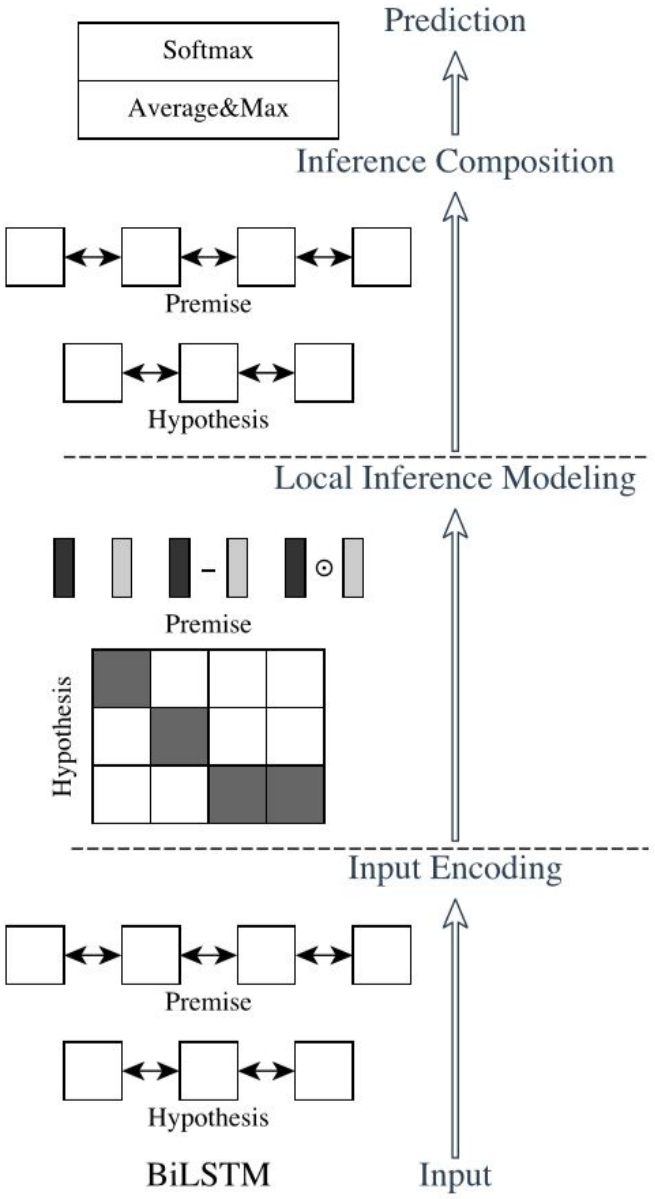
We can use different encoder (here BiLSTM)**:**



$$\bar{\mathbf{a}}_i = \text{BiLSTM}(\mathbf{a}, i), i \in (1, \ldots, \ell_a)$$
$$\bar{\mathbf{b}}_j = \text{BiLSTM}(\mathbf{b}, j), j \in (1, \ldots, \ell_b)$$

where $\bar{\mathbf{a}}_i$ is the output layer of BiLSTM at timestep i of the premise, encoding word a_i and its context. $\bar{\mathbf{b}}_j$ is for the hypothesis.

Enhanced Sequential Inference Models (ESIM)



Layer 3: Inference Composition/Aggregation

Perform composition/aggregation over local inference output to make the global judgement.

Layer 2: Local Inference Modelling

Collect information to perform "local" inference between words or phrases. (Some heuristics works well in this layer.)

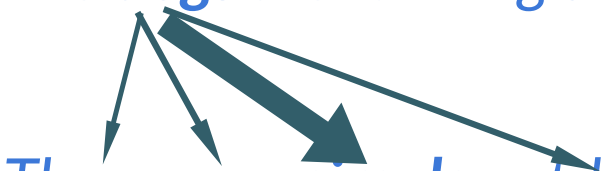
Layer 1: Input Encoding

ESIM uses BiLSTM, but different architectures can be used here, e.g., transformer-based, densely connected CNN, tree-based models, etc.

Local Inference Modelling

Premise

*Two **dogs** are running through a field*



Hypothesis

*There are **animals** outdoors*

Local Inference Modelling

Premise *Two **dogs** are running through a field*

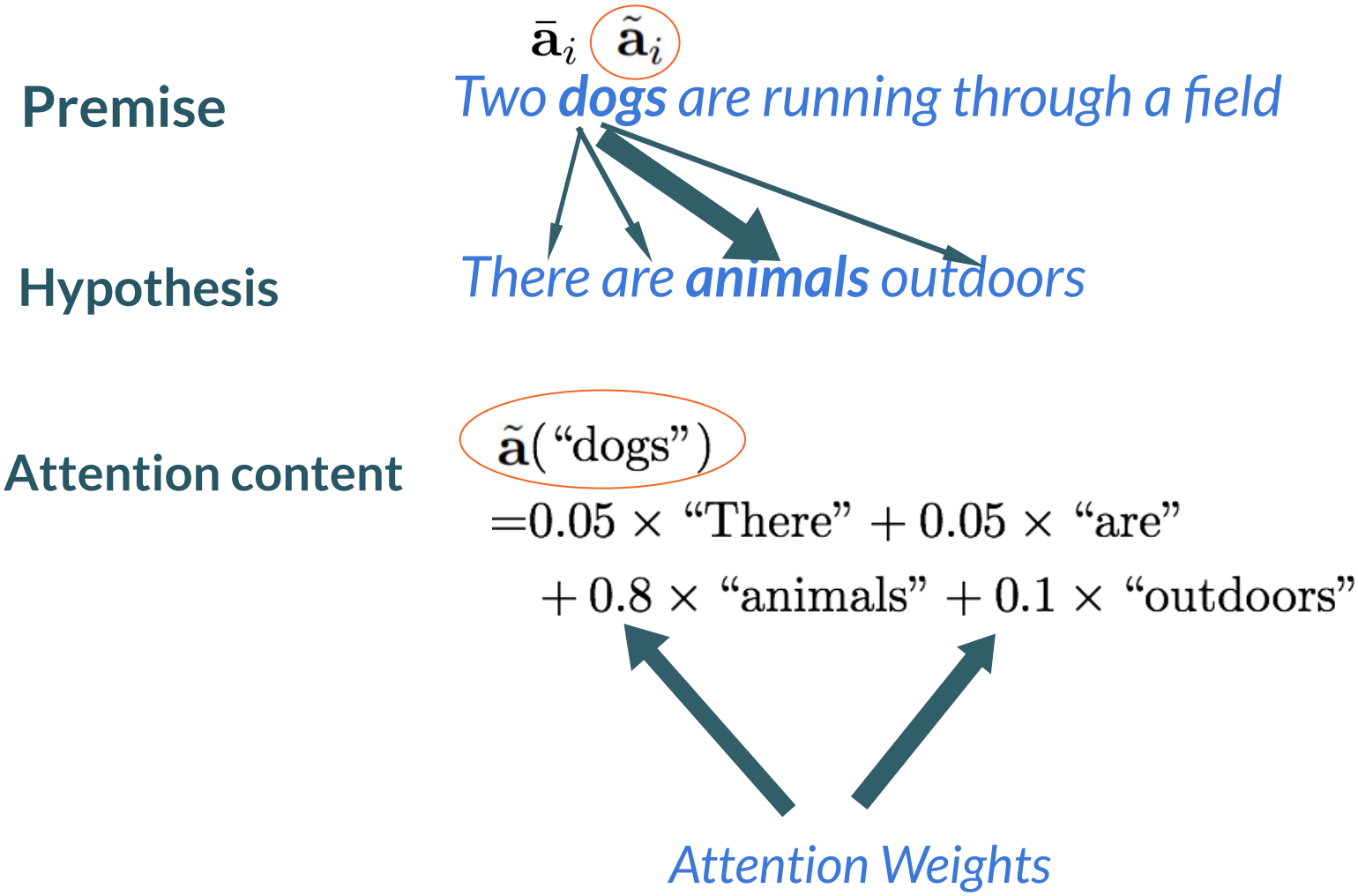
Hypothesis *There are **animals** outdoors*

Attention content

$$\tilde{\mathbf{a}}(\text{“dogs”}) = 0.05 \times \text{“There”} + 0.05 \times \text{“are”} + 0.8 \times \text{“animals”} + 0.1 \times \text{“outdoors”}$$

Attention Weights

Local Inference Modelling



Local Inference Modelling

- The (cross-sentence) *attention content* is computed along both the premise->hypothesis and hypothesis->premise direction.

$$\tilde{\mathbf{a}}_i = \sum_{j=1}^{\ell_b} \frac{\exp(e_{ij})}{\sum_{k=1}^{\ell_b} \exp(e_{ik})} \bar{\mathbf{b}}_j$$

$$\tilde{\mathbf{b}}_j = \sum_{i=1}^{\ell_a} \frac{\exp(e_{ij})}{\sum_{k=1}^{\ell_a} \exp(e_{kj})} \bar{\mathbf{a}}_i$$

where,

$$e_{ij} = \bar{\mathbf{a}}_i^T \bar{\mathbf{b}}_j$$

(more complicated functio $e_{ij} = f(\bar{\mathbf{a}}_i, \bar{\mathbf{b}}_j)$
seemed not further helpful.)

Local Inference Modelling

- With soft alignment ready, we can collect local inference information.
- Note that in various NLI models, the following heuristics have shown to work very well.

$$\mathbf{m}_a = [\bar{\mathbf{a}}; \tilde{\mathbf{a}}; \bar{\mathbf{a}} - \tilde{\mathbf{a}}; \bar{\mathbf{a}} \odot \tilde{\mathbf{a}}]$$
$$\mathbf{m}_b = [\bar{\mathbf{b}}; \tilde{\mathbf{b}}; \bar{\mathbf{b}} - \tilde{\mathbf{b}}; \bar{\mathbf{b}} \odot \tilde{\mathbf{b}}]$$

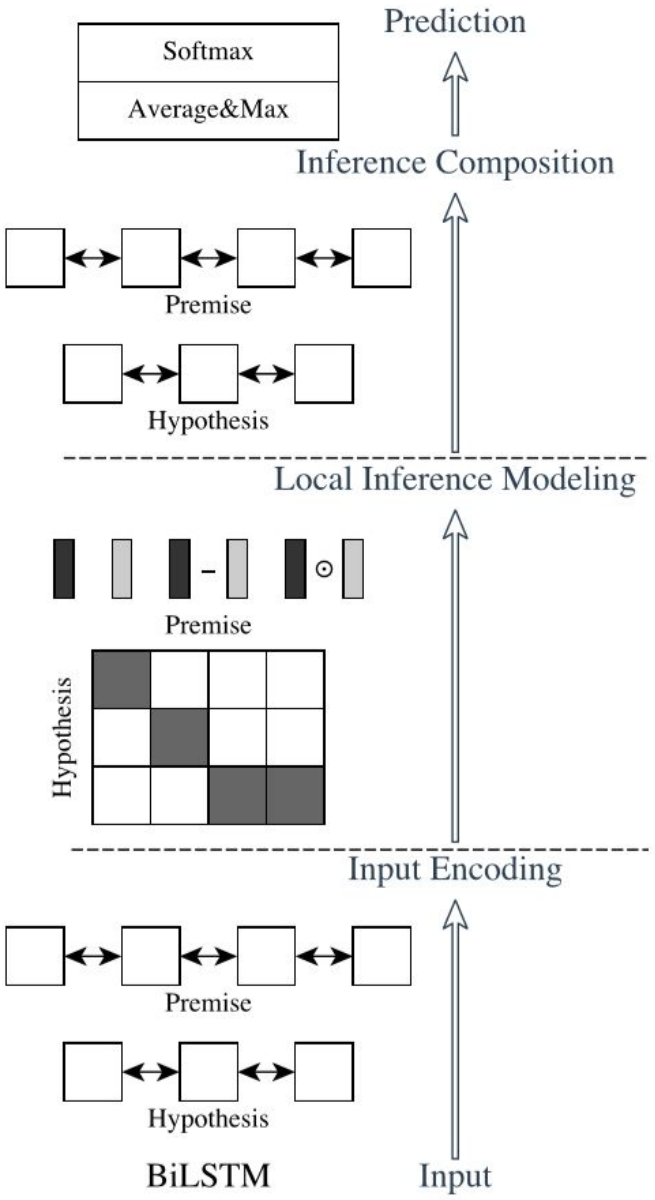
  - Concatenating the two vectors, $\bar{\mathbf{a}}_i$ and $\tilde{\mathbf{a}}_i$ together with their:
    - element-wise product,
    - element-wise difference.

Local Inference Modelling

- Some questions:
  - Instead of using chain RNN, how about other NN architectures like tree-structured?
  - How if one has access to more knowledge than that presents in training data?
    - E.g., lexical entailment information like Minneapolis is part of Minnesota.

  We will come back to these questions later.

Enhanced Sequential Inference Models (ESIM)



Layer 3: Inference Composition/Aggregation

Perform composition/aggregation over local inference output to make the global judgement.

Layer 2: Local Inference Modelling

Collect information to perform "local" inference between words or phrases. (Some heuristics works well in this layer.)

Layer 1: Input Encoding

ESIM uses BiLSTM, but different architectures can be used here, e.g., transformer-based, densely connected CNN, tree-based models, etc.

Inference Composition/Aggregation

- The next component is to perform composition/aggregation over local inference knowledge collected above.
- BiLSTM can be used here to perform "composition" over local inference:

$$\mathbf{v_a} = \text{BiLSTM}(\mathbf{m_a})$$
$$\mathbf{v_b} = \text{BiLSTM}(\mathbf{m_b})$$

where,

$$\mathbf{m}_a = [\bar{\mathbf{a}}; \tilde{\mathbf{a}}; \bar{\mathbf{a}} - \tilde{\mathbf{a}}; \bar{\mathbf{a}} \odot \tilde{\mathbf{a}}]$$
$$\mathbf{m}_b = [\bar{\mathbf{b}}; \tilde{\mathbf{b}}; \bar{\mathbf{b}} - \tilde{\mathbf{b}}; \bar{\mathbf{b}} \odot \tilde{\mathbf{b}}]$$

- Pool with a concatenation of average and max-pooling, and then feed the output to a MLP classifier.

Performance of ESIM on SNLI 🎓

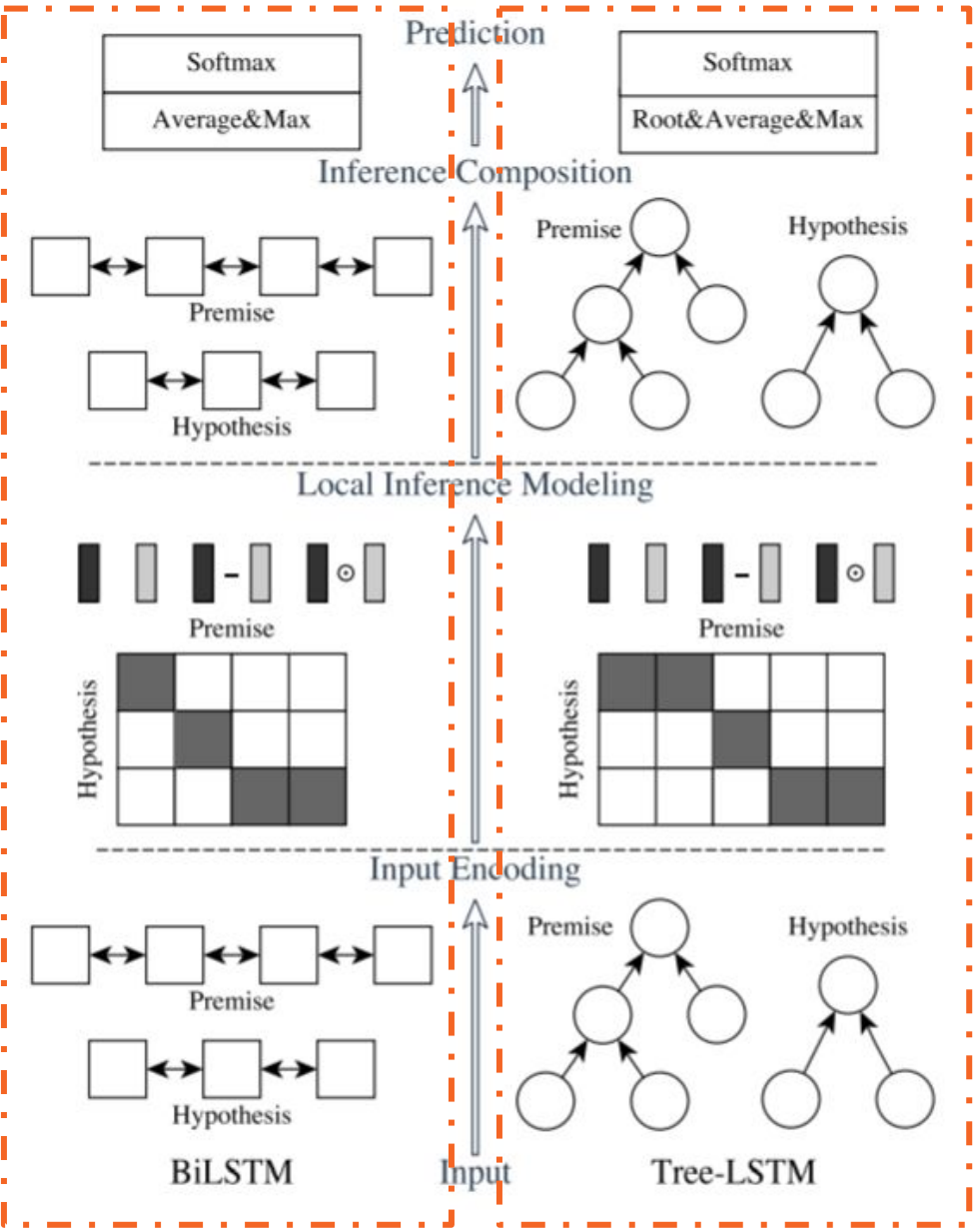
Model	#Para.	Train	Test
(1) Handcrafted features (Bowman et al., 2015)	-	99.7	78.2
(2) 300D LSTM encoders (Bowman et al., 2016)	3.0M	83.9	80.6
(3) 1024D pretrained GRU encoders (Vendrov et al., 2015)	15M	98.8	81.4
(4) 300D tree-based CNN encoders (Mou et al., 2016)	3.5M	83.3	82.1
(5) 300D SPINN-PI encoders (Bowman et al., 2016)	3.7M	89.2	83.2
(6) 600D BiLSTM intra-attention encoders (Liu et al., 2016)	2.8M	84.5	84.2
(7) 300D NSE encoders (Munkhdalai and Yu, 2016a)	3.0M	86.2	84.6
(8) 100D LSTM with attention (Rocktäschel et al., 2015)	250K	85.3	83.5
(9) 300D mLSTM (Wang and Jiang, 2016)	1.9M	92.0	86.1
(10) 450D LSTMN with deep attention fusion (Cheng et al., 2016)	3.4M	88.5	86.3
(11) 200D decomposable attention model (Parikh et al., 2016)	380K	89.5	86.3
(12) Intra-sentence attention + (11) (Parikh et al., 2016)	580K	90.5	86.8
(13) 300D NTI-SLSTM-LSTM (Munkhdalai and Yu, 2016b)	3.2M	88.5	87.3
(14) 300D re-read LSTM (Sha et al., 2016)	2.0M	90.7	87.5
(15) 300D btree-LSTM encoders (Paria et al., 2016)	2.0M	88.6	87.6
(16) 600D ESIM	4.3M	92.6	88.0

Models Enhanced with Syntactic Structures

Models Enhanced with Syntactic Structures

- Syntactic structures have been used in many non-neural NLI/RTE systems (MacCartney, et al., 2013; Dagan et al. 2013).
- How to incorporate linguistic structures into NN based NLI systems? Some typical models:
  - **H**ierarchical **I**nference **M**odels (**HIM**) (Chen et al., 2017)
  - **S**tack-augmented **P**arser-**I**nterpreter **N**eural **N**etwork (SPINN) (Bowman et al., 2016) and follow-up work. (sentence-vector-based models)
  - **T**ree-**B**ased **CNN** (TBCNN) (Mou et al., 2016)

ESIM HIM

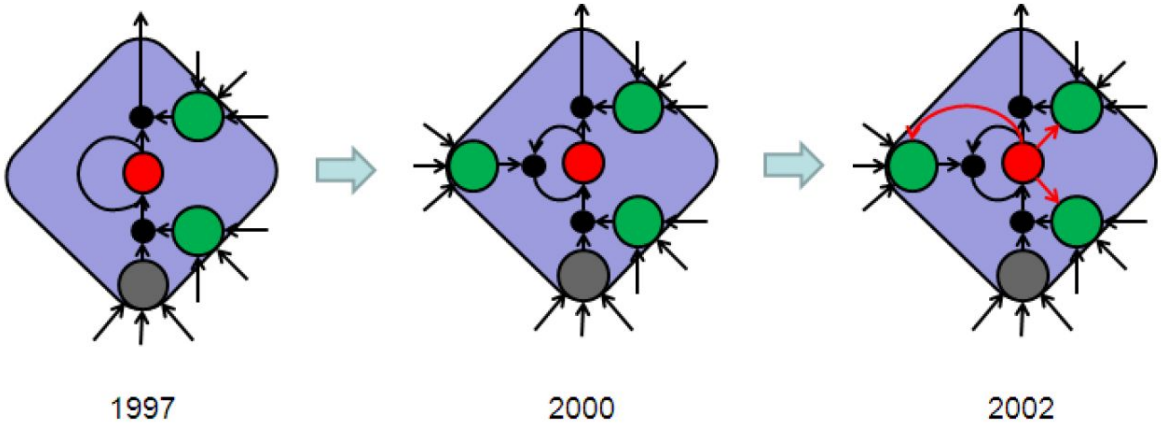


Parse information can be considered in different phases of NLI.

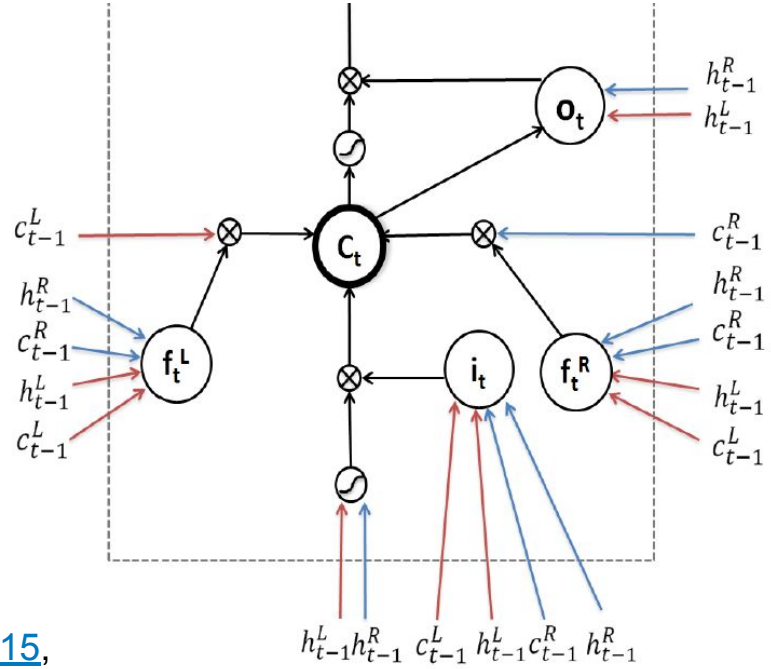
Chen et al. '17

Tree LSTM

Chain LSTM

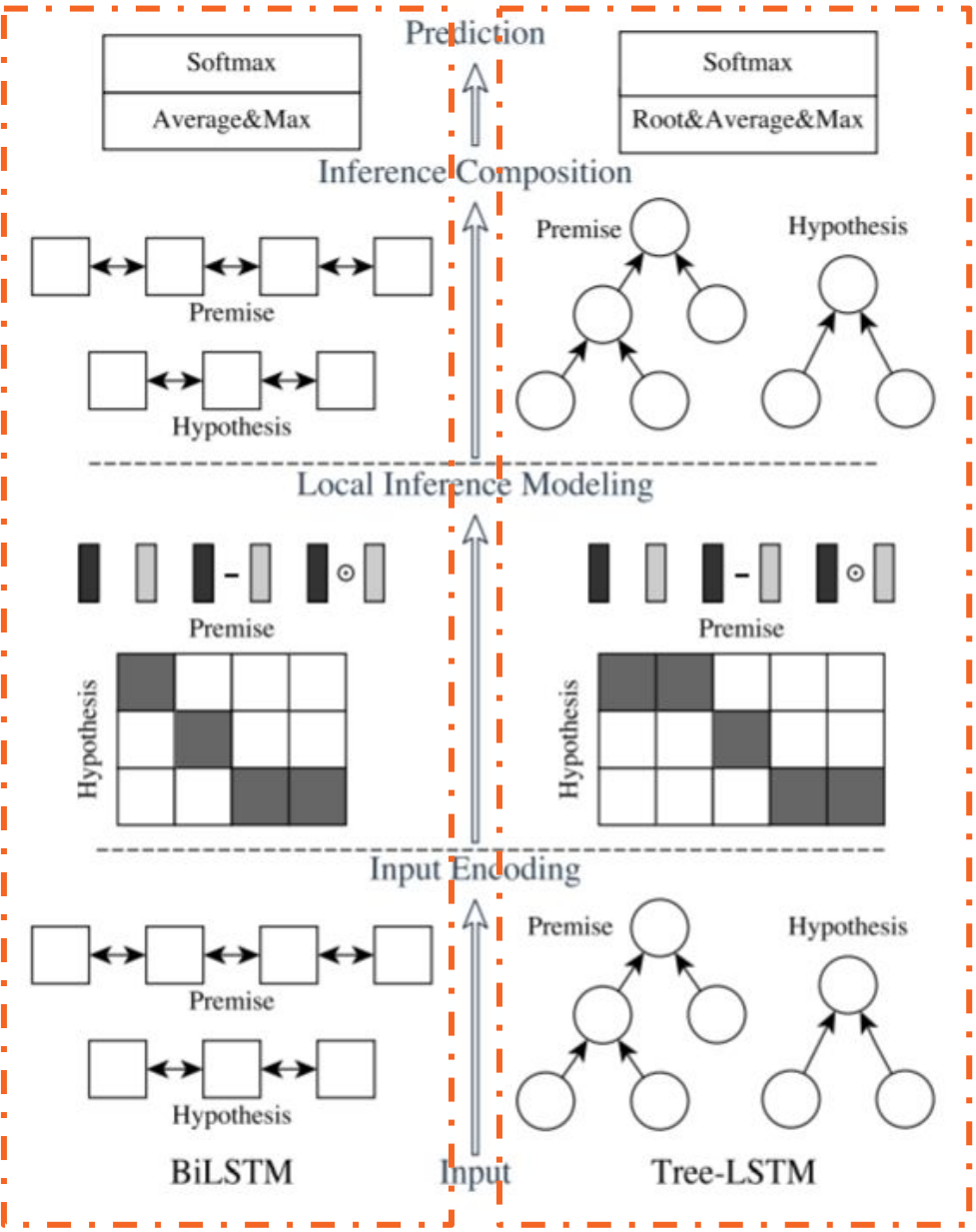


Tree LSTM



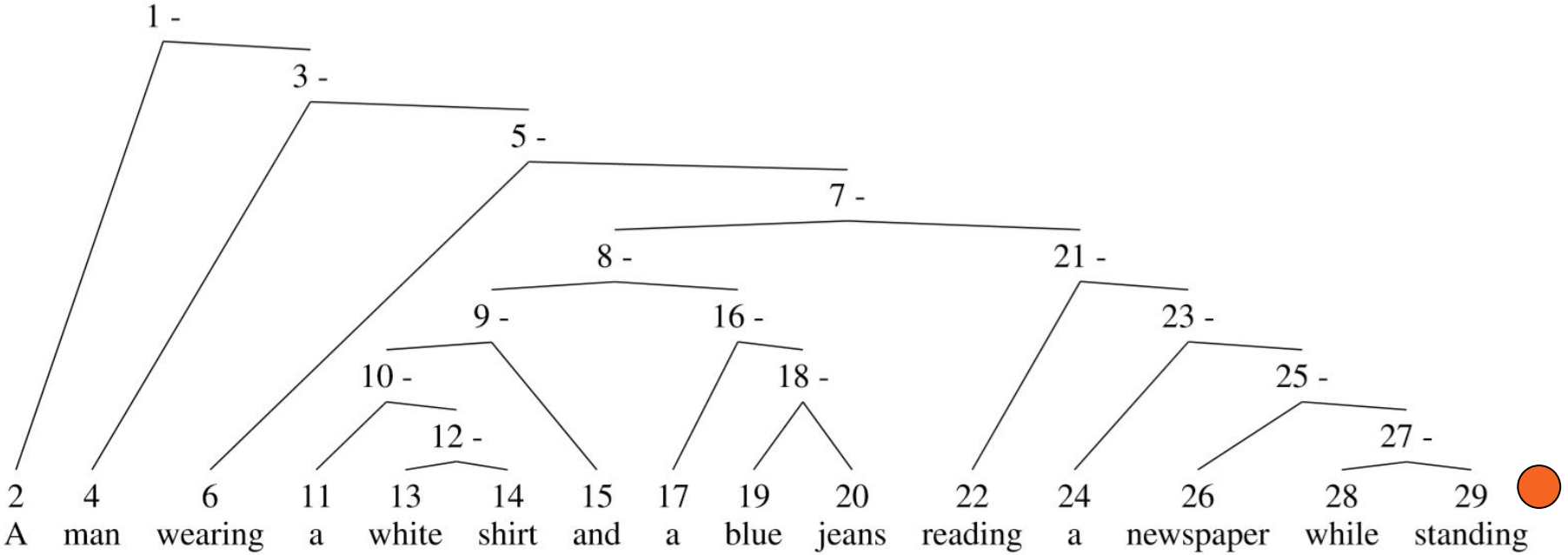
Tai et al. '15, Le et al. '15, Zhu et al. '15,

ESIM HIM

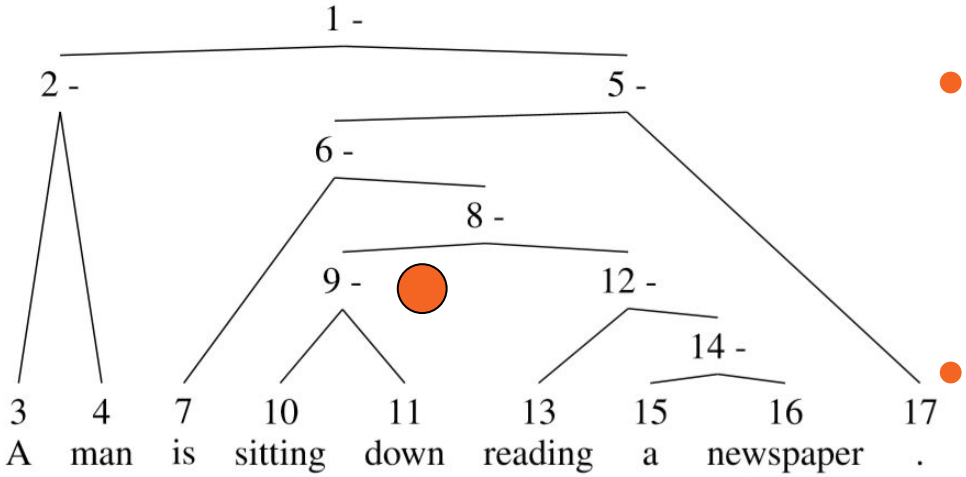


Parse information can be first used to encode input sentences.

Chen et al. '17

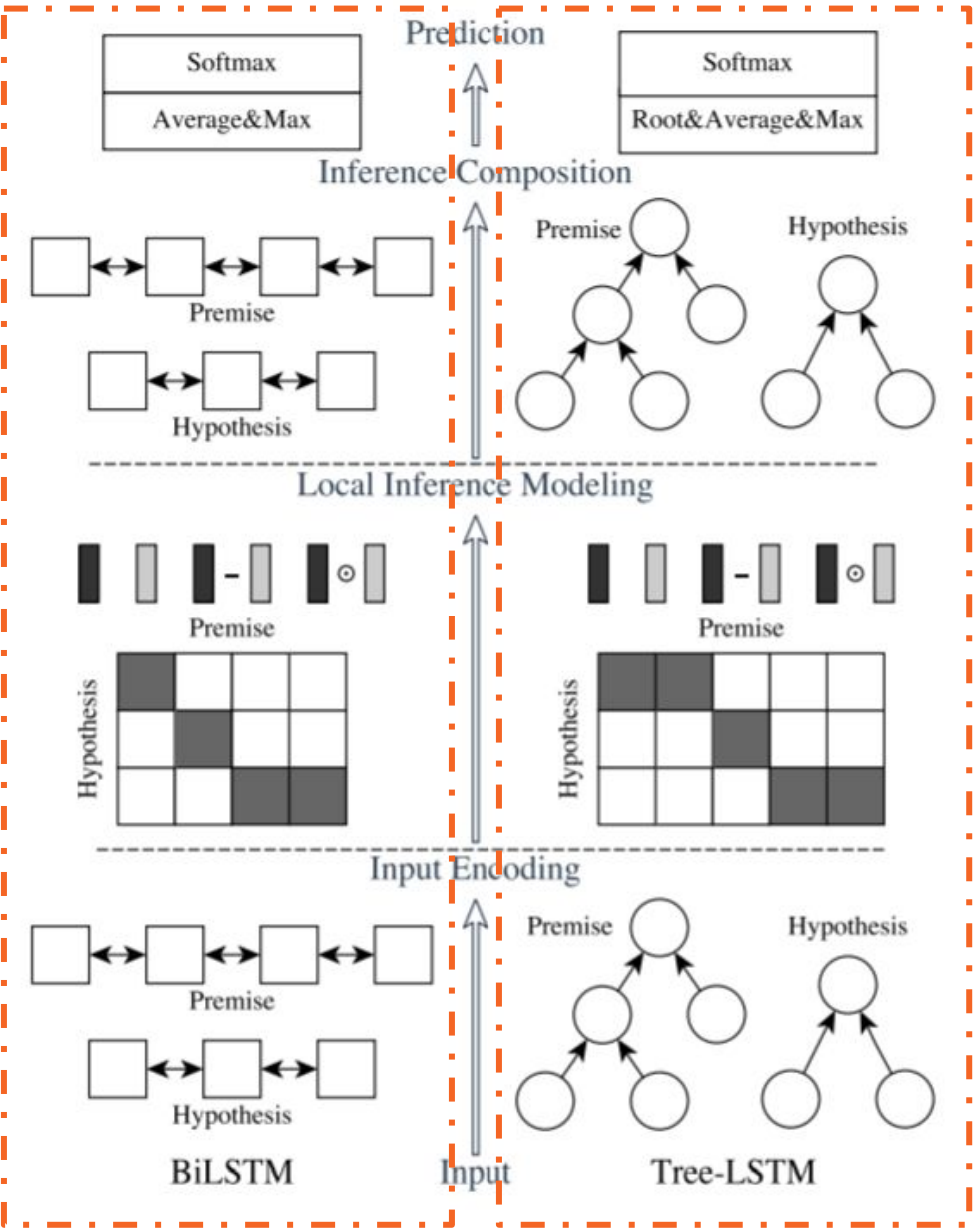


(a) Binarized constituency tree of premise



(b) Binarized constituency tree of hypothesis

- Attention weights showed that the tree models aligned “sitting down” with “standing” and the classifier relies on that to make the right judgement
- The sequential model, however, aligned “sitting” with “reading” and “standing” equally and confused the classifier.



Perform “composition” over local inference information on trees:

$$\mathbf{v}_{a,t} = \mathrm{TrLSTM}(F(\mathbf{m}_{a,t}), \mathbf{h}^L_{t-1}, \mathbf{h}^R_{t-1})$$
$$\mathbf{v}_{b,t} = \mathrm{TrLSTM}(F(\mathbf{m}_{b,t}), \mathbf{h}^L_{t-1}, \mathbf{h}^R_{t-1})$$

where, $m_{a,t}$ and $m_{b,t}$ are passed through a feed-forward layer to reduce the parameter number and alleviate overfitting.

Performance

Model	#Para.	Train	Test
(1) Handcrafted features (Bowman et al., 2015)	-	99.7	78.2
(2) 300D LSTM encoders (Bowman et al., 2016)	3.0M	83.9	80.6
(3) 1024D pretrained GRU encoders (Vendrov et al., 2015)	15M	98.8	81.4
(4) 300D tree-based CNN encoders (Mou et al., 2016)	3.5M	83.3	82.1
(5) 300D SPINN-PI encoders (Bowman et al., 2016)	3.7M	89.2	83.2
(6) 600D BiLSTM intra-attention encoders (Liu et al., 2016)	2.8M	84.5	84.2
(7) 300D NSE encoders (Munkhdalai and Yu, 2016a)	3.0M	86.2	84.6
(8) 100D LSTM with attention (Rocktäschel et al., 2015)	250K	85.3	83.5
(9) 300D mLSTM (Wang and Jiang, 2016)	1.9M	92.0	86.1
(10) 450D LSTMN with deep attention fusion (Cheng et al., 2016)	3.4M	88.5	86.3
(11) 200D decomposable attention model (Parikh et al., 2016)	380K	89.5	86.3
(12) Intra-sentence attention + (11) (Parikh et al., 2016)	580K	90.5	86.8
(13) 300D NTI-SLSTM-LSTM (Munkhdalai and Yu, 2016b)	3.2M	88.5	87.3
(14) 300D re-read LSTM (Sha et al., 2016)	2.0M	90.7	87.5
(15) 300D btree-LSTM encoders (Paria et al., 2016)	2.0M	88.6	87.6
(16) 600D ESIM	4.3M	92.6	88.0
(17) HIM (600D ESIM + 300D Syntactic tree-LSTM)	7.7M	93.5	**88.6**

Effects of Different Components: Alation Analysis

Model	Train	Test
(17) HIM (ESIM + syn.tree)	93.5	88.6
(18) ESIM + tree	91.9	88.2
(16) ESIM	92.6	88.0
(19) ESIM - ave./max	92.9	87.1
(20) ESIM - diff./prod.	91.5	87.0
(21) ESIM - inference BiLSTM	91.3	87.3
(22) ESIM - encoding BiLSTM	88.7	86.3
(23) ESIM - P-based attention	91.6	87.2
(24) ESIM - H-based attention	91.4	86.5
(25) syn.tree	92.9	87.8

Ablation Analysis

Tree Models for Entailment in Formal Logic

- Evans et al (2018) constructed a dataset and studied deep learning models to detect entailment in formal logic, such as:

$$p \models p \vee q \qquad \neg p \wedge \neg q \models \neg q \qquad p \not\models \neg q \qquad \neg p \wedge \neg q \not\models p \vee q$$

- The aim is to help understand two questions:
  - "Can neural networks understand logical formulae well enough to detect entailment?
  - "Which architectures are the best?
- When annotating the data, efforts were made to avoid annotation artifacts.
  - E.g. formulas in positive and negative examples must have the same distribution over length.

Evans et al. '18

Tree Models for Entailment in Formal Logic

	model	valid	test (easy)	test (hard)	test (big)	test (massive)	test (exam)
baselines	Linear BoW	52.6	51.4	50.0	49.7	50.0	52.0
	MLP BoW	57.8	57.1	51.0	55.8	49.9	56.0
benchmark models	Transformer	57.1	56.8	50.8	51.2	50.3	46.9
	ConvNet Encoders	59.3	59.7	52.6	54.9	50.4	54.0
	LSTM Encoders	68.3	68.3	58.1	61.1	52.7	70.0
	BiDirLSTM Encoders	66.6	65.8	58.2	61.5	51.6	78.0
	TreeNet Encoders	72.7	72.2	69.7	67.9	56.6	85.0
	TreeLSTM Encoders	79.1	77.8	74.2	74.2	59.3	75.0
	LSTM Traversal	62.5	61.8	56.2	57.3	50.6	61.0
	BiDirLSTM Traversal	63.3	64.0	55.0	57.9	50.5	66.0
new model	**PossibleWorldNet**	98.7	**98.6**	**96.7**	**93.9**	**73.4**	**96.0**

- The results suggest that, if the structure of input is given, is unambiguous, and is a central feature of the task, models that explicitly exploit structures outperform models which must implicitly model the structure of sequences.

SPINN: Doing Away with Test-Time Tree

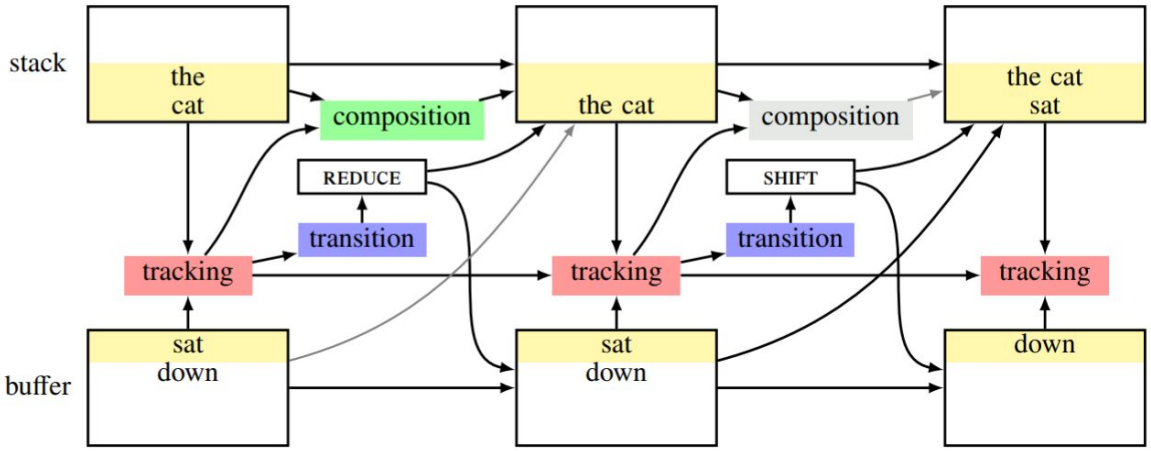


Image credit: Sam Bowman and co-authors.

- Shift-Reduce Parsers:
  - *Shift* unattached leaves from a buffer onto a processing stack.
  - *Reduce* the top two child nodes on the stack to a single parent node.

SPINN: Jointly train a TreeRNN and a vector-based shift-reduce parser.

Training time trees offer supervision for shift-reduce parser.
No need for test time trees!

Bowman et al ‘16

SPINN: Doing Away with Test-Time Tree

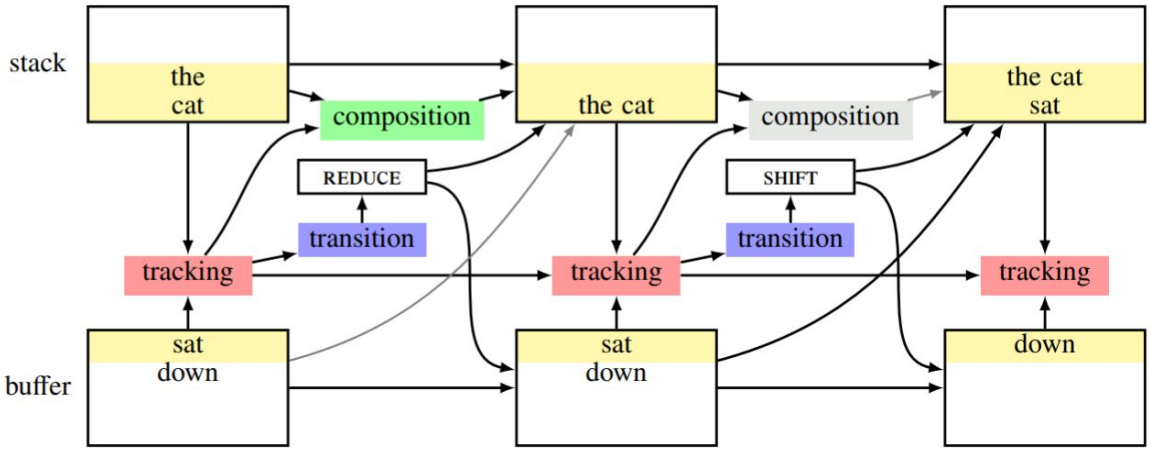


Image credit: Sam Bowman and co-authors.

- Word vectors start on buffer ***b*** (top: first word in sentence).
- *Shift* moves word vectors from buffer to stack ***s***.
- *Reduce* pops top two vectors off the stack, applies $f^R: R^d \times R^d \rightarrow R^d$, and pushes the result back to the stack (i.e. TreeRNN composition).
- *Tracker LSTM* tracks parser/composer state across operations, decides shift-reduce operations and is supervised by both observed shift-reduce operations and end-task.

SPINN + RL: Doing Away with Training-Time Tree

- Identical to SPINN at test time, but uses the reinforce algorithm at training time to compute gradients for the transition classification function.
- Better than LSTM baselines: model captures and exploits structure.
- Model is not biased by what linguists think trees should be like.

Yogatama et al. '17

Do Latent Tree Learning Identify Meaningful Structures?

- Williams et al. (2018) conducted a comprehensive comparison on existing models that use explicit linguistic tree and latent trees.
  - The models include those proposed by Yogatama et al., (2016), Choi et al. (2018), and variants of SPINN models.
- Their main findings are:
  - The learned latent trees are helpful in the construction of semantic representations for sentences.
  - The best available models for latent tree learning learn grammars that do not correspond to the structures of formal syntax and semantics.

Williams et al. '18

Q & A

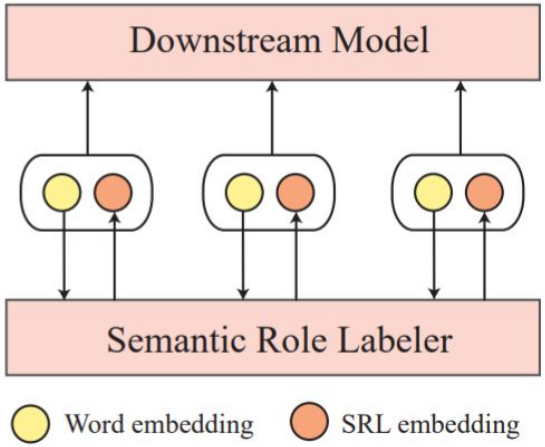
NLI Tutorial

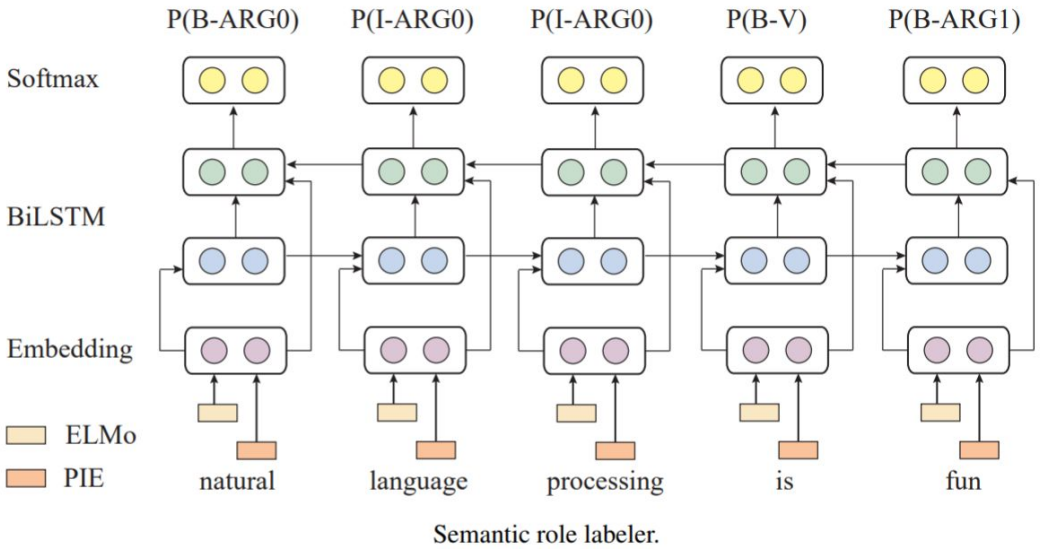
Intermission

Slides: nlitutorial.github.io

Models Enhanced with Semantic Roles

Models Enhanced with Semantic Roles





Semantic role labeler.

- Recent research (Zhang et al., 2019) incorporated SRL into NLI and found it improved various baselines.
- The proposed model simply concatenates SRL embedding into word embedding.

Zhang et al. '19

Models Enhanced with Semantic Roles

- The proposed method is reported to be very effective when used with pretrained models, e.g., ELMo (Peters et al., '17), GPT (Radford et al., 2018), and BERT (Devlin et al., 2018).
  - ELMo: pretrained model is used to initialize an existing NLI model's input-encoding layers. It does not change or replace the NLI model itself.
  - GPT and BERT: pretrained model architectures and parameters are used to perform NLI: parameters will be finetuned in NLI and task-specific networks are no longer used.

Peters et al. '17, Radford et al. '18, Devlin et al. '18

Models Enhanced with Semantic Roles

Model	Accuracy (%)
DIIN	88.0
DR-BiLSTM	88.5
CAFE	88.5
MAN	88.3
KIM	88.6
DMAN	88.8
ESIM + TreeLSTM	88.6
ESIM + ELMo	88.7
DCRCN	88.9
LM-Transformer	89.9
MT-DNN†	91.1
Baseline (ELMo)	88.4
+ SRL	89.1
Baseline ($\text{BERT}_{\text{BASE}}$)	89.2
+ SRL	89.6
Baseline ($\text{BERT}_{\text{LARGE}}$)	90.4
+ SRL	**91.3**

Accuracy on SNLI 

Zhang et al. '19

Modeling External Knowledge

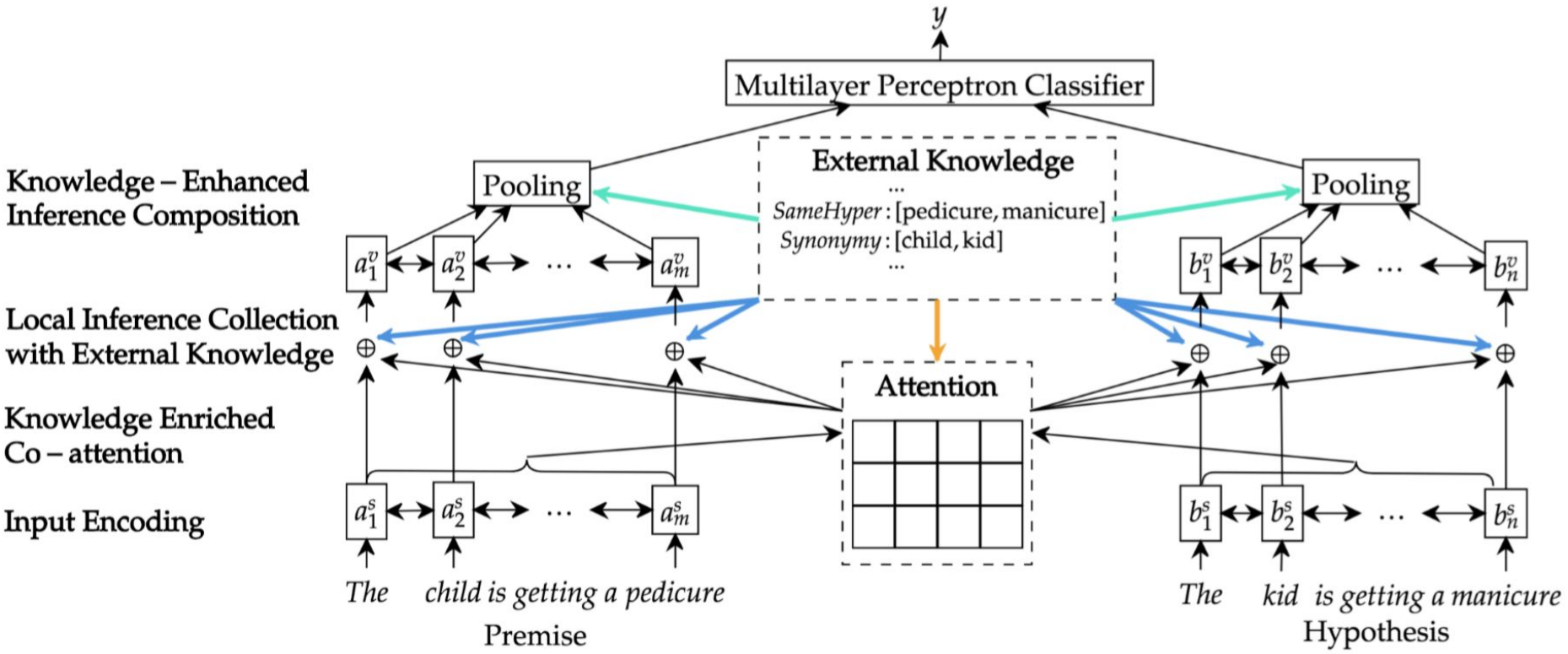
There are two typical ways to add into NLI “external” knowledge that does not present in NLI training data:

- leveraging structured (often human-curated) knowledge,
- using pretrained (self-supervised) models.

Modeling External Knowledge

Leveraging Structured Knowledge

NLI Models Enhanced with External Knowledge



Chen et al. '18

NLI Models Enhanced with External Knowledge

- Knowledge-enhanced co-attention:

$$e_{ij} = (\boldsymbol{a}_i^s)^{\mathrm{T}} \boldsymbol{b}_j^s + F(\boldsymbol{r}_{ij}) .$$

- Local inference with external knowledge:

$$\boldsymbol{a}_i^m = G([\boldsymbol{a}_i^s; \boldsymbol{a}_i^c; \boldsymbol{a}_i^s - \boldsymbol{a}_i^c; \boldsymbol{a}_i^s \odot \boldsymbol{a}_i^c; \sum_{j=1}^{\ell_b} \alpha_{ij} \boldsymbol{r}_{ij}]) ,$$

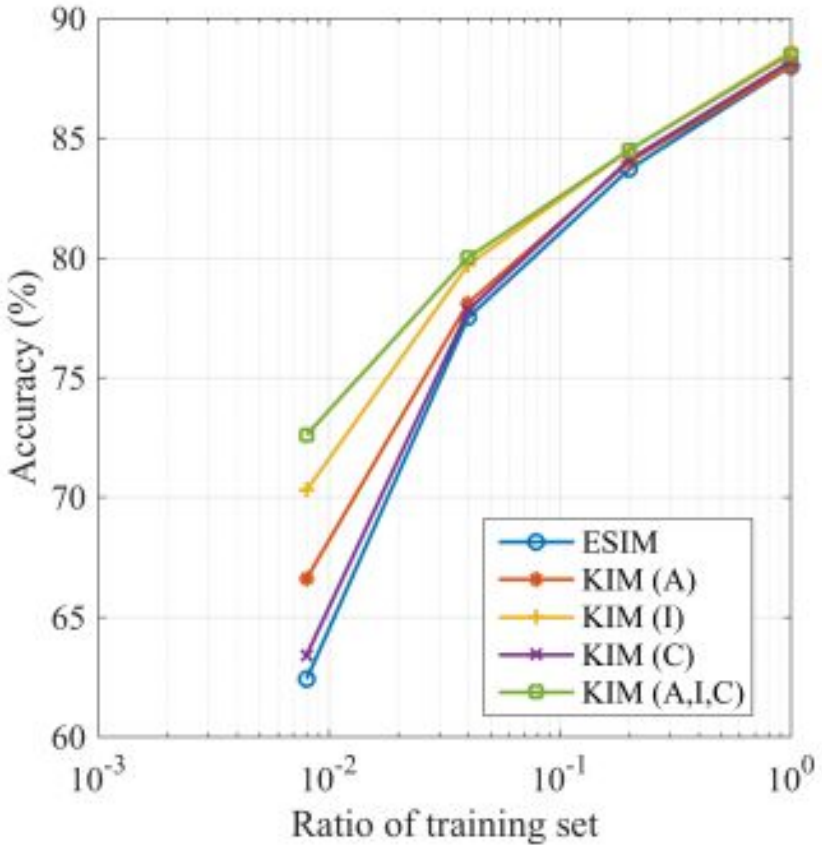
- Enhancing inference composition/aggregation:

$$\boldsymbol{a}^{\mathrm{w}} = \sum_{i=1}^{\ell_a} \frac{\exp(H(\sum_{j=1}^{\ell_b} \alpha_{ij} \boldsymbol{r}_{ij}))}{\sum_{i=1}^{\ell_a} \exp(H(\sum_{j=1}^{\ell_b} \alpha_{ij} \boldsymbol{r}_{ij}))} \boldsymbol{a}_i^v ,$$

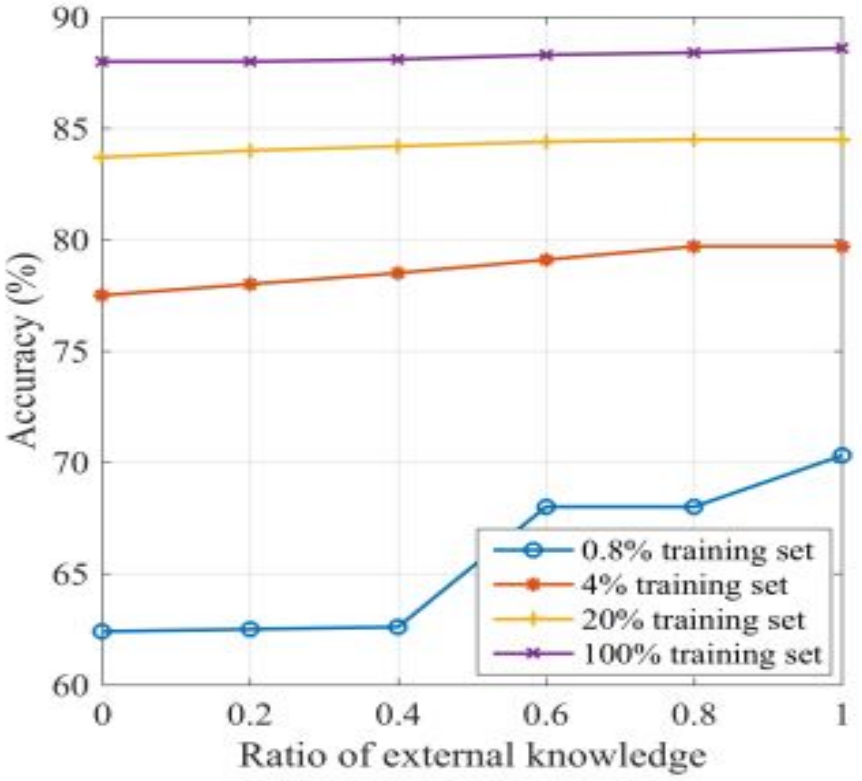
Results on SNLI

Model	Test
LSTM Att. (Rocktäschel et al., 2015)	83.5
DF-LSTMs (Liu et al., 2016a)	84.6
TC-LSTMs (Liu et al., 2016b)	85.1
Match-LSTM (Wang and Jiang, 2016)	86.1
LSTMN (Cheng et al., 2016)	86.3
Decomposable Att. (Parikh et al., 2016)	86.8
NTI (Yu and Munkhdalai, 2017b)	87.3
Re-read LSTM (Sha et al., 2016)	87.5
BiMPM (Wang et al., 2017)	87.5
DIIN (Gong et al., 2017)	88.0
BCN + CoVe (McCann et al., 2017)	88.1
CAFE (Tay et al., 2018)	88.5
ESIM (Chen et al., 2017a)	88.0
KIM	**88.6**

Analysis



Performances of KIM under different sizes of training-data.



Performances of KIM under different amounts of external knowledge.

Results on the Glockner Dataset

Model	SNLI	Glockner's(Δ)
(Parikh et al., 2016)*	84.7	51.9 (-32.8)
(Nie and Bansal, 2017)*	86.0	62.2 (-23.8)
ESIM *	87.9	65.6 (-22.3)
KIM (This paper)	**88.6**	**83.5** (-5.1)

- For a premise in SNLI, Glockner et al. (2018) generated a hypothesis by replacing a single word in premise, in order to test if NLI systems learn simple lexical and word knowledge.

  *A **South** Korean woman gives a manicure.*

  *A **North** North Korean woman gives a manicure.*

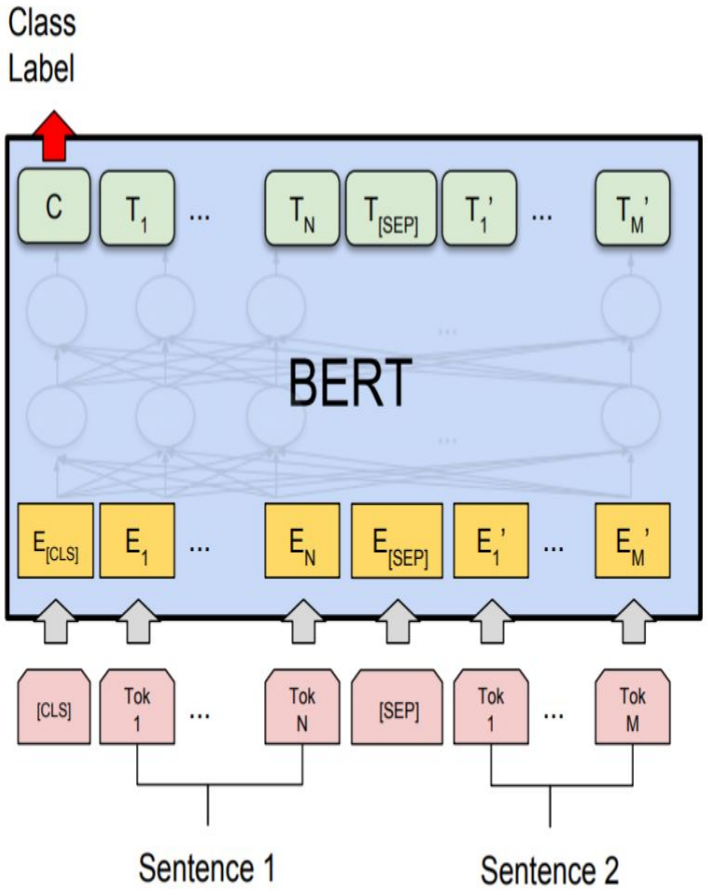
- KIM performs much better than other models on this dataset.

Glockner et al. '18

Modeling External Knowledge

Self-supervised Learning from Unannotated Text

Pretrained Models Leverage Unstructured Data



- The pretrained models can leverage larger unannotated data, which have brought forward the SOTA of NLI, like in many other tasks.
- Bringing other types of external knowledge may further help:
  - We have discussed that Zhang et al. (2019) showed SRL further improve both $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$.
  - We can see more analyses between BERT and KIM.

External Knowledge: BERT vs. KIM

Sentences
P: Yellow banners with a black lion print are hung across some trees in a **sun-lit** neighborhood.
H: Yellow banners with a black lion print are hung across some trees in a **moon-lit** neighborhood.
P: A young boy takes the first step onto **Mars**.
H: A young boy takes the first step onto **Earth**.
P: A Vietnamese woman gives a manicure a **South** Korean woman gives a manicure.
H: A Vietnamese woman gives a manicure a **North** Korean woman gives a manicure.
P: An **Indian** man is perching on top of a wall with a hammer and chisel.
H: An **Indonesian** man is perching on top of a wall with a hammer and chisel.

Examples on which BERT is right but KIM is wrong.

Sentences
P: There are two people **inside**, and two men outside, a cafe; with a tv on in the background.
H: There are two people **outside**, and two men outside, a cafe; with a tv on in the background.

Examples on which KIM is right but BERT is wrong.

Li et al. '19

More Analysis on Pairs of Systems

	BERT	GPT	KIM	ESIM
BERT	.561	.580	**.652**	.616
GPT		.304	.543	.457
KIM			.491	.552
ESIM				.320

- Oracle accuracy of pairs of systems (if one of the two systems under concern makes the correct prediction on a test case, we count it as correct), on a subset of the stress test proposed by Naik et al. (2018).
- The pair, BERT and KIM, complement each other more than other system pairs, e.g., BERT and GPT.

Li et al. '19, Naik et al. '18

Outline

- "Full" deep-learning models for NLI
  - Baseline models and typical components
  - NLI models enhanced with syntactic structures
  - NLI models considering semantic roles and discourse information
  - Incorporating external knowledge
    - Incorporating human-curated structured knowledge
    - Leveraging unstructured data with self-supervision (aka. unsupervised pretraining)
- Sentence-vector-based NLI models
  - A top-ranked model in RepEval-2017
  - Current top models based on dynamic self-attention
- Several additional topics

Sentence-vector-based Models

- As discussed above, NLI is an important test bed for representation learning for sentences

  "*Indeed, a capacity for reliable, robust, open-domain natural language inference is arguably a necessary condition for full natural language understanding (NLU).*" (MacCartney, '09)

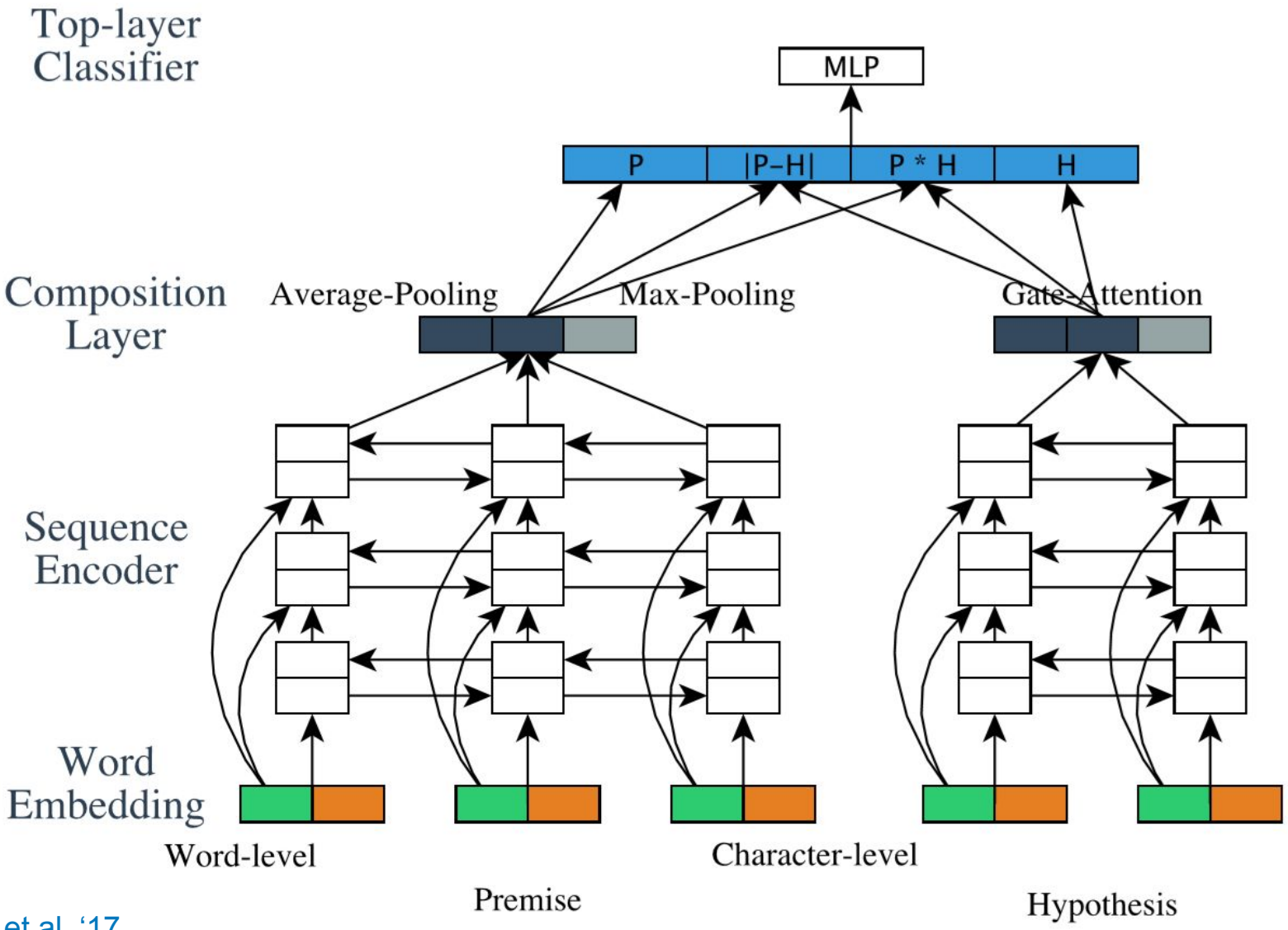
- Sentence-vector-based models encode sentences and test the modeling quality on NLI
  - No cross-sentence attention is allowed, since the goal is to test individual sentence representation.

RepEval-2017 Shared Task

- The RepEval-2017 shared task (Nangia et al. 2017) adopted the MNLI dataset to evaluate sentence representation.
- We will discuss one of the top-ranked models (Chen et al. '17). Other top models can be found in (Nie and Bansal, '17; Balazs et al., '17).

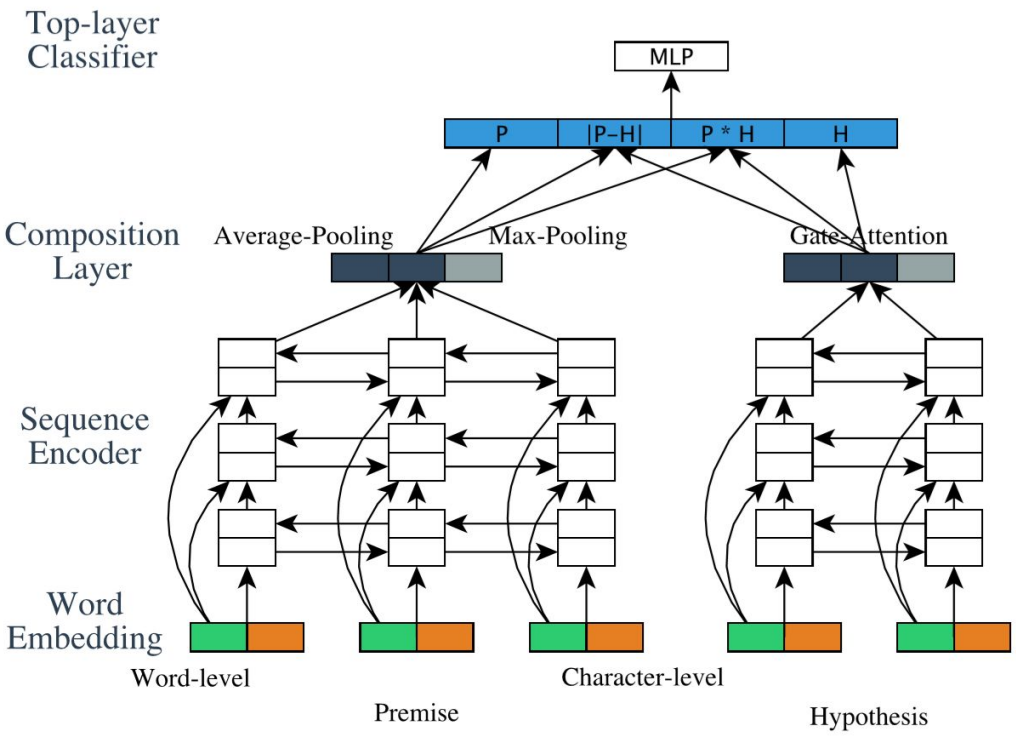
Nangia et al. '17, Nie and Bansal. '17, Conneau et al., '17, Balazs et al., '17, Chen et al. '17

RNN-Based Inference Model with Gated Attention



Chen et al. '17

Gated Attention on Output



- In addition to averaged and max-pooling, weighted average over output is used:

$$v_g^p = \sum_{t=1}^{n} \frac{\|i_t\|_2}{\sum_{j=1}^{n} \|i_j\|_2} h_t^p$$

$$v_g^p = \sum_{t=1}^{n} \frac{\|1 - f_t\|_2}{\sum_{j=1}^{n} \|1 - f_j\|_2} h_t^p$$

$$v_g^p = \sum_{t=1}^{n} \frac{\|o_t\|_2}{\sum_{j=1}^{n} \|o_j\|_2} h_t^p$$

The weights are taken from the input, forget, and output gates of the top-layer BiLSTM.

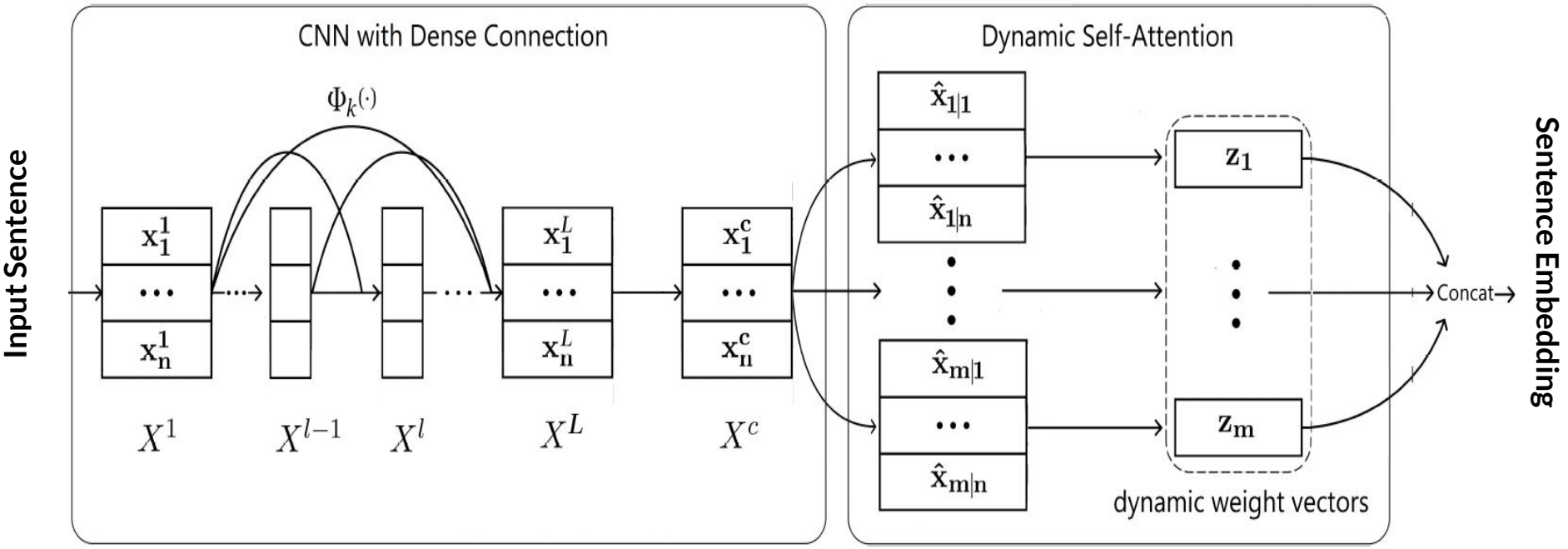
Results

Model	In-Domain	Cross-Domain
CBOW	64.8	64.5
BiLSTM	66.9	66.9
ESIM	72.3	72.1
TALP-UPC*	67.9	68.2
LCT-MALTA*	70.7	70.8
Rivercorners*	72.1	72.1
Rivercorners (ensemble)*	72.2	72.8
YixinNie-UNC-NLP*	74.5	73.5
Single*	73.5	73.6
Ensembled*	74.9	74.9
Single (Input Gate)*	73.5	73.6
Single (Forget Gate)	72.9	73.1
Single (Output Gate)	73.7	73.4
Single - Gated-Att	72.8	73.6
Single - CharCNN	72.9	73.5
Single - Word Embedding	65.6	66.0
Single - AbsDiff/Product	69.7	69.2

Accuracy of models on the MNLI test sets.

Sentence-vector-based models seem to be sensitive to operations performed at the top layer of the networks, e.g., pooling or element-wise diff/product. See (Chen et al, '18) for more work on generalized pooling.

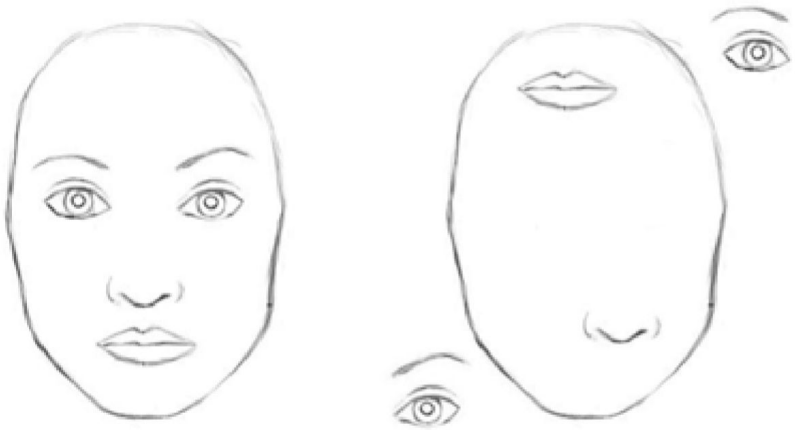
CNN with Dynamic Self-Attention



- So far, this model achieves the best performance on SNLI among sentence-vector-based models.
- Key idea: stack a dynamic self-attention over CNN (with dense connection)
- The proposed dynamic self-attention borrows the ideas from the Capsule Network (Sabour et al. 2017; Hinton et al., 2018).

Yoon et al. '18, Sabour et al. '17, Hinton et al . '18

Capsule Networks

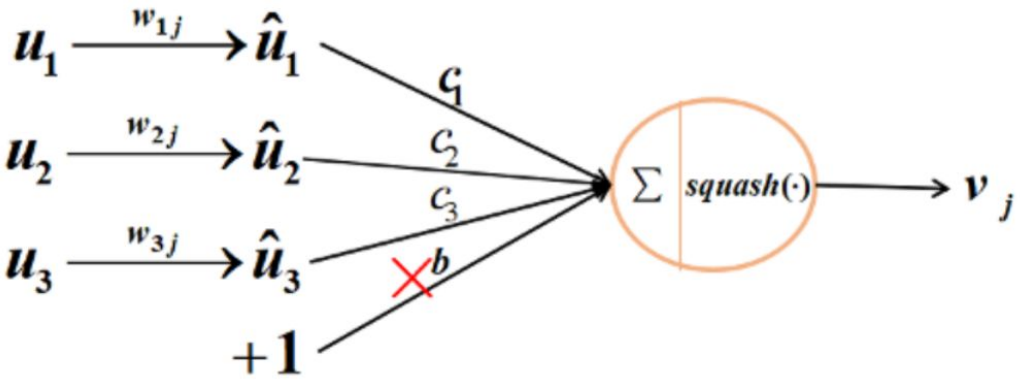


- One important motivation for the Capsule Network is to better model *part-whole* relationship in images.
  - To recognize the left figure is a face but not the right one, the *parts* (here, nose, eyes and mouth) need to agree on how a face should look like (e.g., its position and orientation).
  - Each *part* and the *whole* (here, a face) is represented as a vector.
  - Agreement is computed through *dynamic routing*.

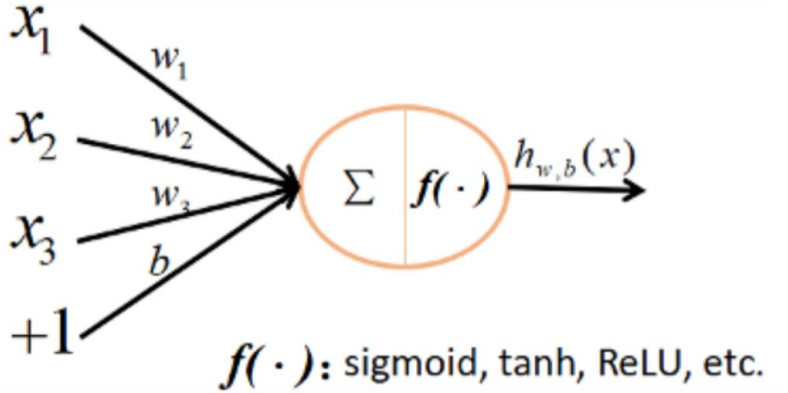
Sabour et al. '17, Hinton et al . '18

Capsule Networks

Capsule cell



Regular neuron



- Key differences:
  - Input of a capsule cell is a number of vectors ($\boldsymbol{u}_1$ is a vector) but not a scalar ($\boldsymbol{x}_1$ is a scalar).
  - Voting parameters c_1, c_2, c_3 are not part of model parameters but learned through *dynamic routing* and are not kept after training.

Sabour et al. '17, Hinton et al . '18

Dynamic Routing

Procedure 1 Routing algorithm.

1: **procedure** ROUTING($\hat{\boldsymbol{u}}_{j|i}, r, l$)
2: for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow 0$.
3: **for** r iterations **do**
4: for all capsule i in layer l: $\mathbf{c}_i \leftarrow \texttt{softmax}(\mathbf{b}_i)$
5: for all capsule j in layer $(l+1)$: $\mathbf{s}_j \leftarrow \sum_i c_{ij} \hat{\mathbf{u}}_{j|i}$
6: for all capsule j in layer $(l+1)$: $\mathbf{v}_j \leftarrow \texttt{squash}(\mathbf{s}_j)$
7: for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow b_{ij} + \hat{\mathbf{u}}_{j|i}.\mathbf{v}_j$
return $\mathbf{v}_j$

- Key ideas:
  - A capsule at a lower layer needs to decide how to send its message to higher level capsules.
  - The essence of the above algorithm is to ensure a lower level capsule will send more message to the higher level capsule that "agrees" with it (indicated by a high similarity between them).

Sabour et al. '17, Hinton et al . '18

CNN with Dynamic Self-Attention for NLI

Algorithm 1 Dynamic Self-Attention

procedure ATTENTION($\hat{\mathbf{x}}_{\mathbf{j|i}}, r$)
  for all i^{th} word, j^{th} attention : $q_{ij} = 0$
  for r iterations **do**
    for all $i, j : a_{ij} = \frac{exp(q_{ij})}{\sum_k exp(q_{kj})}$
    for all $j : \mathbf{s_j} = \sum_i a_{ij} \hat{\mathbf{x}}_{\mathbf{j|i}}$
    for all $j : \mathbf{z_j} = \mathrm{Tanh}(\mathbf{s_j})$
    for all $i, j : q_{ij} = q_{ij} + \hat{\mathbf{x}}^{\mathbf{T}}_{\mathbf{j|i}} \mathbf{z_j}$
  return all $\mathbf{z_j}$

- The proposed model borrows the idea of weight adaptation method in dynamic routing to adapt attention weight (a_{ij}).
- In addition, instead of performing multihead attention, the work performs multiple dynamic self-attention (DSA).

Yoon et al. '18

CNN with Dynamic Self-Attention for NLI

Publications	Model Description	Accuracy
Seonhoon Kim et al. '18	Densely-Connected Recurrent and Co-Attentive	*86.5*
Talman et al. '18	600D Hierarchical BiLSTM with Max Pooling	*86.6*
Qian Chen et al. '18	600D BiLSTM with generalized pooling	*86.6*
Kiela et al. '18	512D Dynamic Meta-Embeddings	*86.7*
Deunsol Yoon et al. '18	600D Dynamic Self-Attention Model	*86.8*
Deunsol Yoon et al. '18	2400D Multiple-Dynamic Self-Attention Model	*87.4*

Current leaderboard of sentence-vector-based models on SNLI (as of June 1st, 2019).

Revisiting Artifacts of Data

Breaking NLI Systems with Sentences that Require Simple Lexical Inferences

- As discussed above, Glockner et al. (2018) create a new test set that shows the deficiency of NLI systems in modelling lexical and world knowledge.
- The set is developed upon the SNLI's test set: for a premise sentence, a hypothesis is constructed by replacing one word in premise.

Premise/Hypothesis	Label
The man is holding a saxophone The man is holding an electric guitar	contradiction[1]
A little girl is very sad. A little girl is very unhappy.	entailment
A couple drinking wine A couple drinking champagne	neutral

Glockner et al. '18

Breaking NLI Systems with Sentences that Require Simple Lexical Inferences

- The performance on the new test set is substantially worse across systems, suggesting some drawback of the existing NLI datasets.

Model	Train set	SNLI test set	New test set	Δ
Decomposable Attention (Parikh et al., 2016)	SNLI	84.7%	51.9%	-32.8
	MultiNLI + SNLI	84.9%	65.8%	-19.1
	SciTail + SNLI	85.0%	49.0%	-36.0
ESIM (Chen et al., 2017)	SNLI	87.9%	65.6%	-22.3
	MultiNLI + SNLI	86.3%	74.9%	-11.4
	SciTail + SNLI	88.3%	67.7%	-20.6
Residual-Stacked-Encoder (Nie and Bansal, 2017)	SNLI	86.0%	62.2%	-23.8
	MultiNLI + SNLI	84.6%	68.2%	-16.8
	SciTail + SNLI	85.0%	60.1%	-24.9
WordNet Baseline	-	-	85.8%	-
KIM (Chen et al., 2018)	SNLI	88.6%	83.5%	-5.1

Accuracy of models on SNLI and the Glockner dataset.

“Stress Tests” for NLI

- Naik et al. (2018) propose an evaluation methodology consisting of automatically constructed test examples.
- The “stress tests” constructed are organized into three classes:
  - Competence test: numerical reasoning and antonymy understanding.
  - Distraction test: robustness on lexical similarity, negation, word overlap.
  - Noise test: robustness on “spelling error”.

Naik et al. ‘18

“Stress Tests” for NLI

System	Original MultiNLI Dev		Competence Test			Distraction Test						Noise Test	
			Antonymy		Numerical Reasoning	Word Overlap		Negation		Length Mismatch		Spelling Error	
	Mat	Mis	Mat	Mis		Mat	Mis	Mat	Mis	Mat	Mis	Mat	Mis
NB	74.2	74.8	15.1	19.3	21.2	47.2	47.1	39.5	40.0	48.2	47.3	51.1	49.8
CH	73.7	72.8	11.6	9.3	30.3	58.3	58.4	52.4	52.2	63.7	65.0	68.3	69.1
RC	71.3	71.6	36.4	32.8	30.2	53.7	54.4	49.5	50.4	48.6	49.6	66.6	67.0
IS	70.3	70.6	14.4	10.2	28.8	50.0	50.2	46.8	46.6	58.7	59.4	58.3	59.4
BiLSTM	70.2	70.8	13.2	9.8	31.3	57.0	58.5	51.4	51.9	49.7	51.2	65.0	65.1
CBOW	63.5	64.2	6.3	3.6	30.3	53.6	55.6	43.7	44.2	48.0	49.3	60.3	60.6

Classification accuracy (%) of state-of-the-art models on the stress tests. Three of the models, **NB** (Nie and Bansal, ‘17), **CH** (Chen et al., ‘17), and **RC** (Balazs et al., ‘17) are models submitted to RepEvel-2017. **IS** (Conneau et al., 2017) a model proposed to learn general sentence embedding trained on NLI.

Nie and Bansal. ‘17, Conneau et al., ‘17, Balazs et al., ‘17, Chen et al. ‘17

Swapping Premise and Hypothesis

- Wang et al. (2018) proposed the following idea: swapping the premise and hypothesis in the test set to create the diagnostic test.
- For entailment, a better model is supposed to report a larger difference of performance on the original test set and swapped test set.
- It should report comparable accuracy between the original test set and swapped test set for contradiction and neural.

Wang et al. ‘18

Swapping Premise and Hypothesis

Model	Label	Dev	Swap-Dev	Diff-Dev	Test	Swap-Test	Diff-Test
CBOW	E	0.877	0.134	0.743	0.856	0.080	0.776
	C	0.706	0.583	0.123	0.740	0.580	0.160
	N	0.874	0.613	0.261	0.659	0.589	0.070
InferSent	E	0.850	0.090	0.760	0.880	0.087	0.793
	C	0.853	0.666	0.187	0.859	0.682	0.177
	N	0.795	0.713	0.082	0.795	0.712	0.083
DGA	E	0.822	0.376	0.446	0.854	0.422	0.432
	C	0.720	0.660	0.060	0.711	0.650	0.061
	N	0.700	0.648	0.052	0.700	0.619	0.081
ESIM	E	0.891	0.301	0.590	0.884	0.324	0.560
	C	0.865	0.702	0.163	0.861	0.701	0.160
	N	0.806	0.721	0.085	0.801	0.720	0.081
KIM	E	0.908	0.103	0.805	0.895	0.095	0.800
	C	0.850	0.772	0.078	0.845	0.796	0.049
	N	0.800	0.664	0.136	0.781	0.675	0.106
ADV	E	0.862	0.856	0.006	0.854	0.860	-0.006
	C	0.753	0.643	0.110	0.751	0.646	0.105
	N	0.706	0.509	0.197	0.705	0.507	0.198

Performance (accuracy) of different models on the original and swapped SNLI test set. Bigger differences (Diff-Test) for entailment (label E) suggests better models for entailment. Models that consider external semantic knowledge, e.g., KIM, seem to perform better in this swapping test.

More work on analyzing the properties of NLI data: Poliak et. al, '18, Talman and Chatzikyriakidis.

Bringing Explanation to NLI

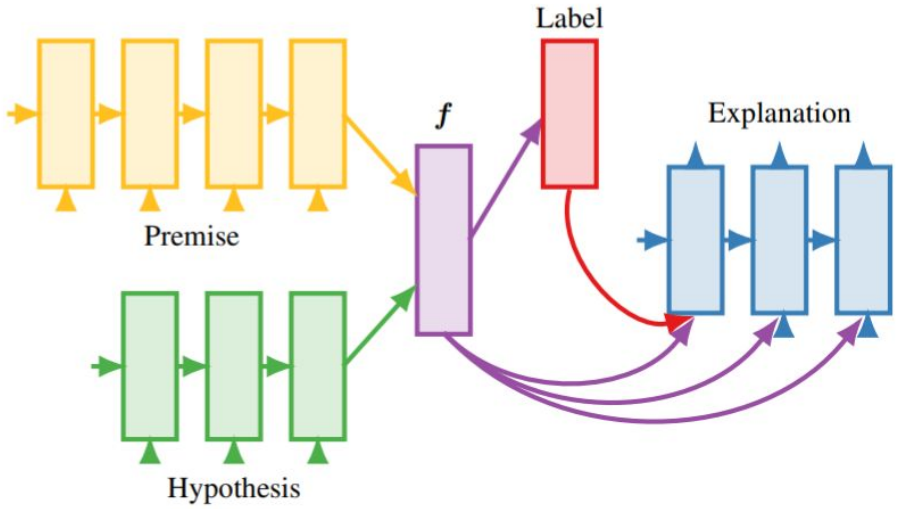
e-SNLI: Bringing Explanation to NLI

Premise: An adult dressed in black holds a stick. Hypothesis: An adult is walking away, empty-handed. Label: contradiction Explanation: Holds a stick implies using hands so it is not empty-handed.
Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her. Hypothesis: A young mother is playing with her daughter in a swing. Label: neutral Explanation: Child does not imply daughter and woman does not imply mother.
Premise: A man in an orange vest leans over a pickup truck. Hypothesis: A man is touching a truck. Label: entailment Explanation: Man leans over a pickup truck implies that he is touching it.

- Extend the SNLI dataset with an additional layer of human-annotated natural language explanation of the entailment relations.
- More research problems can be further explored:
  - Not just predict a label but also generate explanation
  - Obtain full sentence justifications of a model's decision
  - Help transfer to out-of-domain NLI task

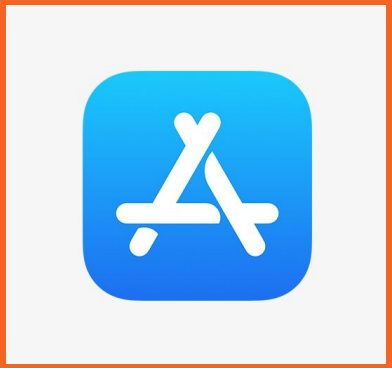
Camburu et al. '18

e-SNLI: Bringing Explanation to NLI



Overview of the e-INFERSENT architecture.

- PREMISEAGNOSTIC: Generate an explanation given only the hypothesis
- PREDICTANDEXPLAIN: Jointly predict a label and generate an explanation for the predicted label
- EXPLAINTHENPREDICT: Generate an explanation then predict a label
- REPRESENT: Universal sentence representations
- TRANSFER: Transfer without fine-tuning to out-of-domain NLI



Natural Language Inference: **Applications**

Applications

Three major application types for NLI:

- *Direct application* of trained NLI models.
- NLI as a *research and evaluation* task for new methods.
- NLI as a *pretraining* task in transfer learning.

Direct Applications

2018 Fact Extraction and Verification shared task (**FEVER**):

Inspired by issues surrounding fake news and automatic fact checking:

"The task challenged participants to classify whether human-written factoid claims could be SUPPORTED or REFUTED using evidence retrieved from Wikipedia"

Thorne et al. '18, Nie et al. '18

Direct Applications

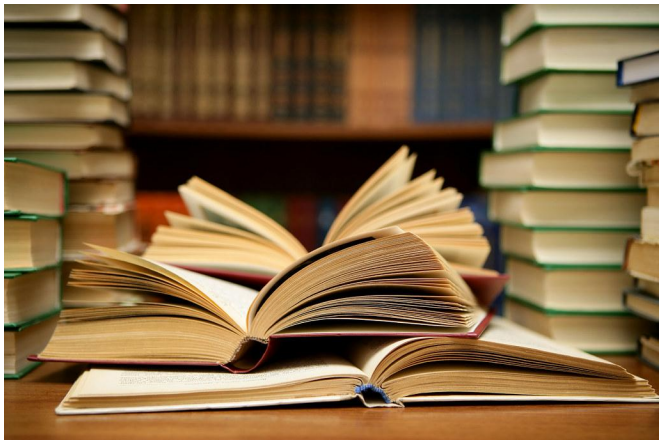
2018 Fact Extraction and Verification shared task (**FEVER**):

Inspired by issues surrounding fake news and automatic fact checking.

SNLI/MNLI models used in many systems, including winner, to decide whether a piece of evidence supports a claim.

Thorne et al. ‘18, Nie et al. ‘18

Direct Applications

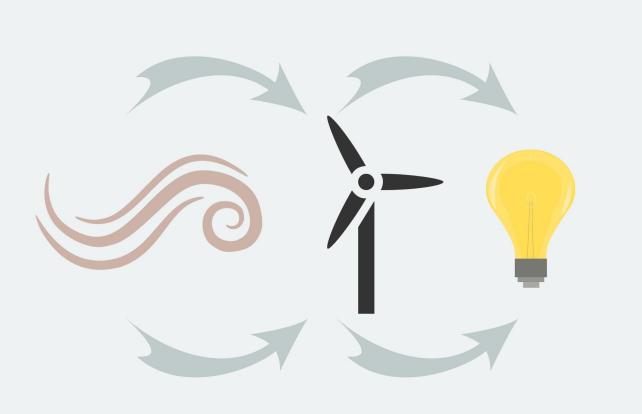


Multi-hop reading comprehension tasks like MultiRC or OpenBook require models to answer a question by combining multiple pieces of evidence from some long text.

Integrating an **SNLI/MNLI-trained ESIM model** into a larger model in two places helps to select and combine relevant evidence for a question.

Trivedi et al. ‘19 (NAACL)

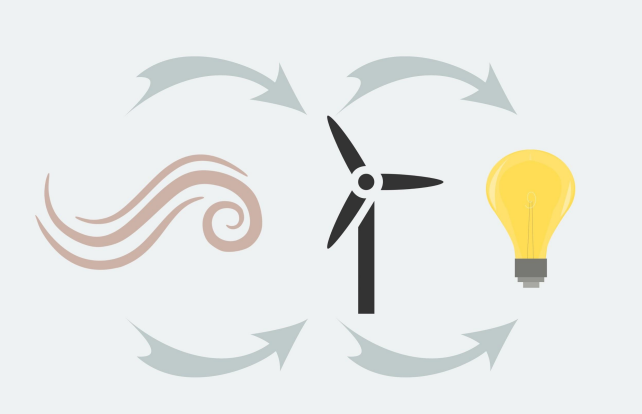
Direct Applications



When **generating video captions**, using an SNLI/MNLI-trained entailment model as part of the objective function can lead to more effective training.

Pasunuru and Bansal ‘17

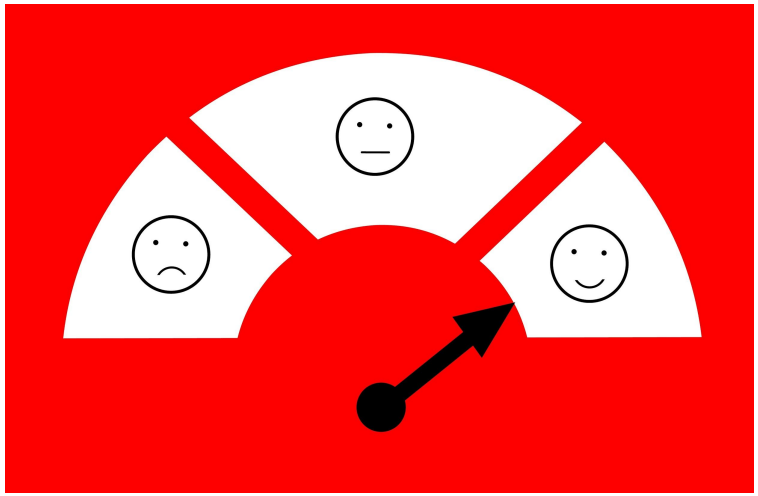
Direct Applications



When **generating long-form text**, using an SNLI/MNLI-trained entailment model as a *cooperative discriminator* can prevent a language model from contradicting itself.

Holtzman et al. ‘18

Evaluation

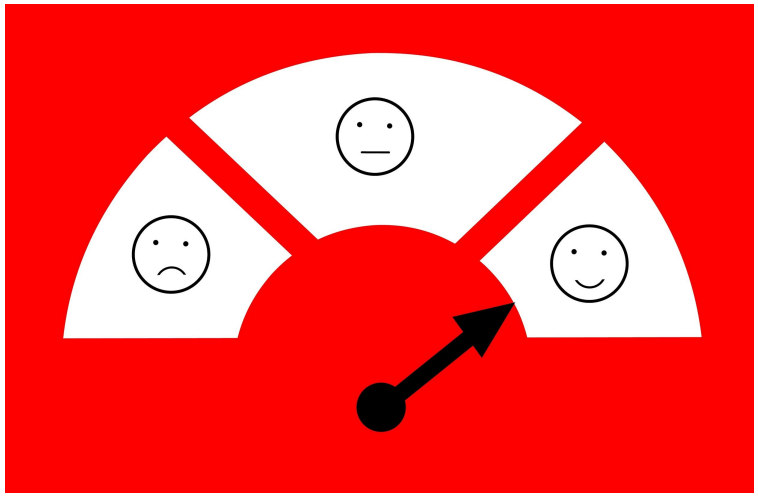


Several entailment corpora have become **established benchmark datasets for studying new ML methods in NLP**.

Used as a major evaluation when developing self-attention networks, language model pretraining, and much more.

Rocktäschel et al. 16, Parikh et al. ‘17, Peters et al. ‘18, Devlin et al. ‘19 (NAACL)

Evaluation



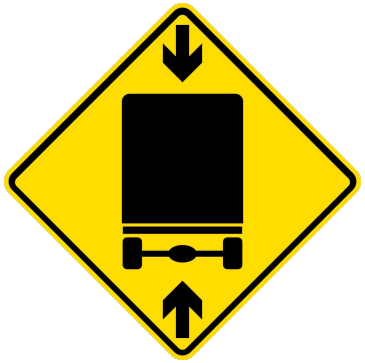
Several entailment corpora have become **established benchmark datasets for studying new ML methods in NLP**.

Used as a major evaluation when developing self-attention networks, language model pretraining, and much more.

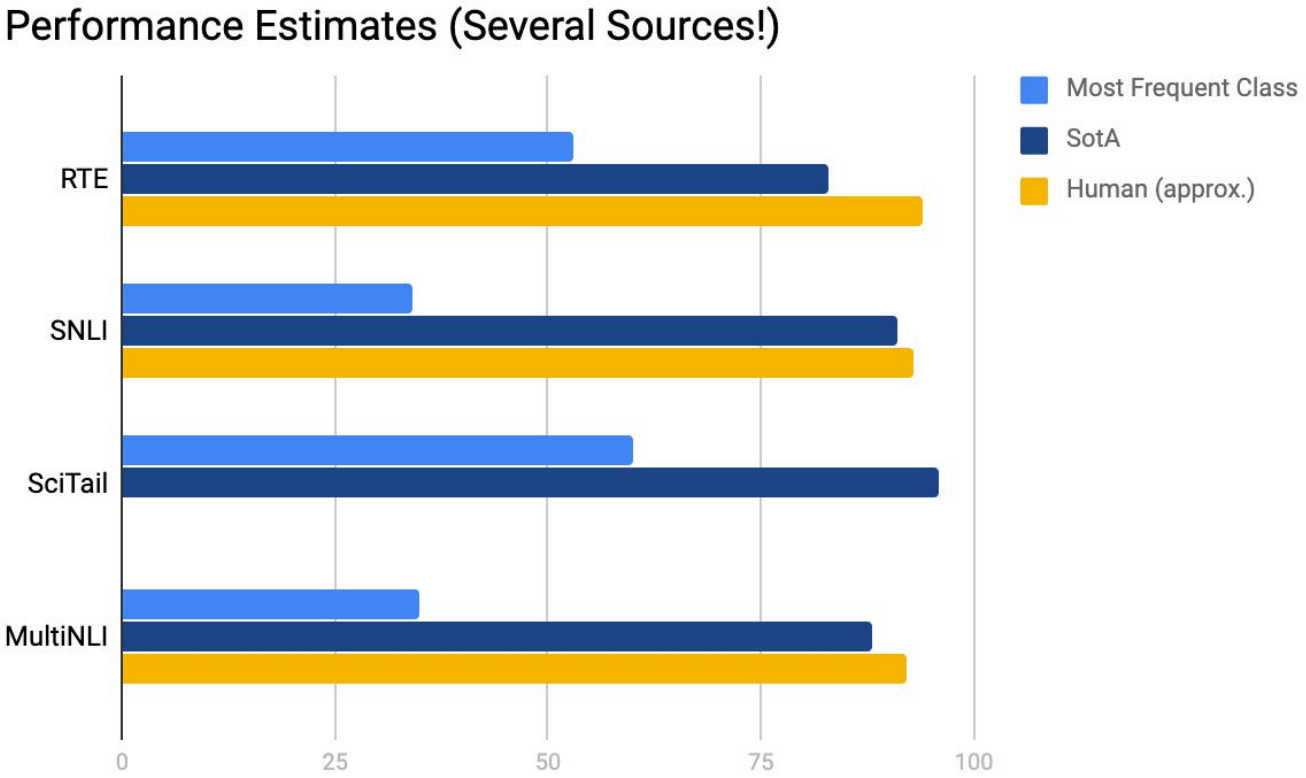
Also included in the SentEval, GLUE, DecaNLP, and SuperGLUE benchmarks and associated software toolkits.

Rocktäschel et al. 16, Parikh et al. ‘17, Peters et al. ‘18, Devlin et al. ‘19 (NAACL)

Evaluation (a Caveat)



State of the art models are **very close to human performance** on major evaluation sets:



Transfer Learning

Training neural network models on large NLI datasets (especially MNLI) and then fine-tuning them on target tasks often yields substantial improvements in target task performance.

Conneau et al. ‘17, Subramanian et al. ‘18, Phang et al. ‘18, Liu et al. ‘19

Transfer Learning

Training neural network models on large NLI datasets (especially MNLI) and then fine-tuning them on target tasks often yields substantial improvements in target task performance.

This works well even in conjunction with strong baselines for pretraining like SkipThought, ELMo, or BERT.

Responsible for the current state of the art on the GLUE benchmark.

Conneau et al. '17, Subramanian et al. '18, Phang et al. '18, Liu et al. '19

Summary and Conclusions

Summary

- The tutorial covers the recent advance on NLI (aka. RTE) research, which is powered by:
  - Large annotated datasets
  - Deep learning models over distributed representation
- We view and discuss NLI as an important test bed for representation learning for natural language.
- We discuss the existing and potential applications of NLI.

Future Work

- Better supervised models (of course)
- Harder naturalistic benchmark datasets
- Explainability
- Better Unsupervised DL approaches
- Application of NLI on more NLP tasks
- Multimodal NLI
- NLI in domains: adaptation
- ...

Thanks!

Questions?

Slides and contact information:
nlitutorial.github.io

Extra Slides

XNLI: Evaluating Cross-lingual Sentence Representations

- As NLI is a good test bed for NLU, cross-lingual NLI can be a good test bed for cross-lingual NLU.
- XNL: cross-lingual NLI dataset for 15 languages, each having 7,500 NLI sentence pairs and in total 112,500 pairs.
  - Following the the construction processing used to construct the MNLI corpora.
- Can be used to evaluate both cross-lingual NLI model and multilingual text embedding model.

Conneau et al. ‘18

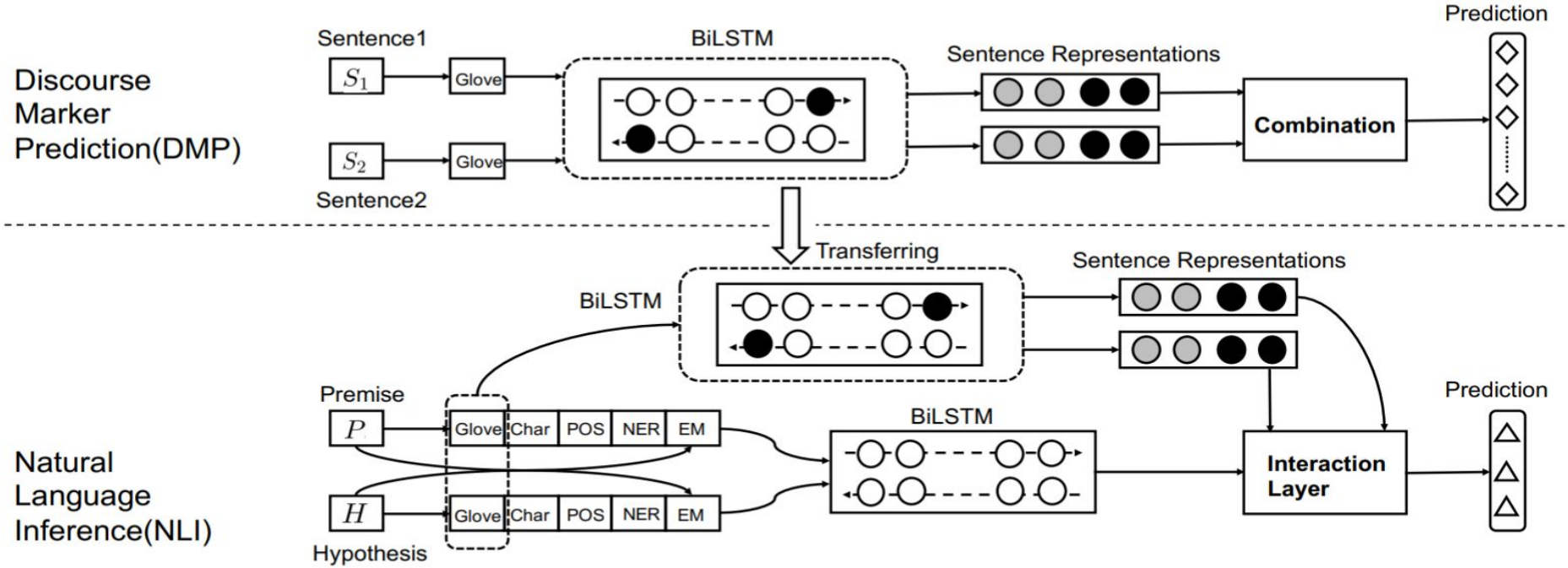
XNLI: Evaluating Cross-lingual Sentence Representations

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
Machine translation baselines (TRANSLATE TRAIN)															
BiLSTM-last	71.0	66.7	67.0	65.7	65.3	65.6	65.1	61.9	63.9	63.1	61.3	65.7	61.3	55.2	55.2
BiLSTM-max	**73.7**	68.3	68.8	66.5	66.4	67.4	66.5	64.5	65.8	66.0	62.8	67.0	62.1	58.2	56.6
Machine translation baselines (TRANSLATE TEST)															
BiLSTM-last	71.0	68.3	68.7	66.9	67.3	68.1	66.2	64.9	65.8	64.3	63.2	66.5	61.8	60.1	58.1
BiLSTM-max	**73.7**	**70.4**	**70.7**	**68.7**	**69.1**	**70.4**	**67.8**	**66.3**	**66.8**	**66.5**	**64.4**	**68.3**	**64.2**	**61.8**	**59.3**
Evaluation of XNLI multilingual sentence encoders (in-domain)															
X-BiLSTM-last	71.0	65.2	67.8	66.6	66.3	65.7	63.7	64.2	62.7	65.6	62.7	63.7	62.8	54.1	56.4
X-BiLSTM-max	**73.7**	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4
Evaluation of pretrained multilingual sentence encoders (transfer learning)															
X-CBOW	64.5	60.3	60.7	61.0	60.5	60.4	57.8	58.7	57.5	58.8	56.9	58.8	56.3	50.4	52.2

Test accuracy of baseline models.
See more recent advance in (Lample and Conneau, 2019)

Conneau et al. '18, Lample and Conneau et al. '19

Models Enhanced with Discourse Markers



- The **D**iscourse **M**arker **A**ugmented **N**etwork (DMAN, Pan et al., 2018) uses discourse marker information to guide NLI decision.
  - Inductive bias is built in for discourse-related words like *but*, *although*, *so*, *because*, etc.
  - The **D**iscourse **M**arker **Prediction** (Nie et al., 2017) is incorporated into DMAN through a reinforcement learning component.

Pan et al. '18

Encoding Syntax

- As discussed before, in (Chen et al '17), tree-LSTM is used to leverage constituency parse for both sentence encoding and inference composition.
- Here for time efficiency, a sequentialized parse tree is used (through pre-order traversal of the tree).