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"

"Hypothesis" or "Sentence B"

i'm not sure what the overnight low was
{entails, contradicts, neither}

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

Dagan et al. '05, MacCartney '09
Example from MNLI
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.)
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?
The Big Question

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

Why not?
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?
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

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

*Isobutylphenylpropionic acid is a medicine for headaches.*

{entails, contradicts, neither}?

*Isobutylphenylpropionic acid is a medicine.*

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.

- *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

Dagan et al. '06 et seq.

- 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
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. \( \Rightarrow \) 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
Premise: A boat sank in the Pacific Ocean.

Hypothesis: A boat sank in the Atlantic Ocean.

Label: neutral

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

One event or two? Two.
One event or two? One, always.

Premise: A boat sank in the Pacific Ocean.

Hypothesis: A boat sank in the Atlantic Ocean.

Label: contradiction
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
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.
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

50
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.
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

<table>
<thead>
<tr>
<th>Tag</th>
<th>SNLI</th>
<th>MultiNLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronouns (PTB)</td>
<td>34</td>
<td>68</td>
</tr>
<tr>
<td>Quantifiers</td>
<td>33</td>
<td>63</td>
</tr>
<tr>
<td>Modals (PTB)</td>
<td>&lt;1</td>
<td>28</td>
</tr>
<tr>
<td>Negation (PTB)</td>
<td>5</td>
<td>31</td>
</tr>
<tr>
<td>‘Wh’ Words (PTB)</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>Belief Verbs</td>
<td>&lt;1</td>
<td>19</td>
</tr>
<tr>
<td>Time Terms</td>
<td>19</td>
<td>36</td>
</tr>
<tr>
<td>Conversational Pivots</td>
<td>&lt;1</td>
<td>14</td>
</tr>
<tr>
<td>Presupposition Triggers</td>
<td>8</td>
<td>22</td>
</tr>
<tr>
<td>Comparatives/Superlatives (PTB)</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>Conditionals</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>Tense Match (PTB)</td>
<td>62</td>
<td>69</td>
</tr>
<tr>
<td>Interjections (PTB)</td>
<td>&lt;1</td>
<td>5</td>
</tr>
<tr>
<td>&gt;20 Words</td>
<td>&lt;1</td>
<td>5</td>
</tr>
<tr>
<td>Existentials (PTB)</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>
The Train-Test Split
## The MNLI Corpus

<table>
<thead>
<tr>
<th>Genre</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Captions (SNLI Corpus)</td>
<td>(550,152)</td>
<td>(10,000)</td>
<td>(10,000)</td>
</tr>
<tr>
<td>Fiction</td>
<td>77,348</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>Government</td>
<td>77,350</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>Slate</td>
<td>77,306</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>Switchboard (Telephone Speech)</td>
<td>83,348</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>Travel Guides</td>
<td>77,350</td>
<td>2,000</td>
<td>2,000</td>
</tr>
</tbody>
</table>
## The MNLI Corpus

<table>
<thead>
<tr>
<th>Genre</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Captions (SNLI Corpus)</td>
<td>(550,152)</td>
<td>(10,000)</td>
<td>(10,000)</td>
</tr>
<tr>
<td>Fiction</td>
<td>77,348</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>Government</td>
<td>77,350</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>Slate</td>
<td>77,306</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>Switchboard (Telephone Speech)</td>
<td>83,348</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>Travel Guides</td>
<td>77,350</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>9/11 Report</td>
<td>0</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>Face-to-Face Speech</td>
<td>0</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>Letters</td>
<td>0</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>OUP (Nonfiction Books)</td>
<td>0</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>Verbatim (Magazine)</td>
<td>0</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>392,702</strong></td>
<td><strong>20,000</strong></td>
<td><strong>20,000</strong></td>
</tr>
</tbody>
</table>
# The MNLI Corpus

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

Good news:

*Most models perform similarly on both sets!*
Annotation Artifacts
Annotation Artifacts

For SNLI:

\( P: \) ???

\( H: \) Someone is not crossing the road.

\textbf{Label:} entailment, contradiction, neutral?

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

Annotation Artifacts


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
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.
In Depth: Natural Logic
Monotonicity

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

- A reptile moved
- A turtle moved
- A reptile danced

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

- No reptile moved
- No turtle moved
- No reptile danced

**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

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X \equiv Y$</td>
<td>equivalence</td>
<td>couch $\equiv$ sofa</td>
</tr>
<tr>
<td>$X \sqsubset Y$</td>
<td>forward entailment</td>
<td>crow $\sqsubset$ bird</td>
</tr>
<tr>
<td>$X \sqsupset Y$</td>
<td>reverse entailment</td>
<td>European $\sqsupset$ French</td>
</tr>
<tr>
<td>$X ^ Y$</td>
<td>negation</td>
<td>human $^$ non-human</td>
</tr>
<tr>
<td>$X \mid Y$</td>
<td>alternation</td>
<td>cat $\mid$ dog</td>
</tr>
<tr>
<td>$X \subseteq Y$</td>
<td>cover</td>
<td>animal $\subseteq$ non-human</td>
</tr>
<tr>
<td>$X # Y$</td>
<td>independence</td>
<td>hungry $#$ hippo</td>
</tr>
</tbody>
</table>
## Beyond Up and Down: Projectivity

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>not-X # not-Y</th>
<th>X ⊢ not-Y</th>
<th>not-X ⊲ Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>X = Y</td>
<td>not-X = not-Y</td>
<td>X</td>
<td>not-Y</td>
<td>not-X</td>
</tr>
<tr>
<td>X # Y</td>
<td>not-X # not-Y</td>
<td>X # not-Y</td>
<td>not-X # Y</td>
<td></td>
</tr>
<tr>
<td>X ⊢ Y</td>
<td>not-X ⊢ not-Y</td>
<td>X</td>
<td>not-Y</td>
<td>not-X # Y</td>
</tr>
<tr>
<td>X ⊲ Y</td>
<td>not-X ⊲ not-Y</td>
<td>X # not-Y</td>
<td>not-X</td>
<td>Y</td>
</tr>
</tbody>
</table>
Chains of Relations

If we know \( A \lor B \) and \( B \land C \), what do we know?

\[
\begin{array}{c}
| \quad \land \quad = \quad \sqsubset \\
\end{array}
\]

So \( A \sqsubseteq C \)

MacCartney and Manning ‘09
Putting it all together

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.*

<table>
<thead>
<tr>
<th>i</th>
<th>$x_i$</th>
<th>$e_i$</th>
<th>$\beta(e_i)$</th>
<th>$\beta(x_{i-1}, e_i)$</th>
<th>$\beta(x_0, x_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Stimpy is a cat</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Stimpy is a dog</td>
<td>$\text{SUB}(\text{cat, dog})$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Stimpy is not a <strong>dog</strong></td>
<td>$\text{INS}(\text{not})$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Stimpy is not a <strong>poodle</strong></td>
<td>$\text{SUB}(\text{dog, poodle})$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
Before we delve into Deep Learning (DL) models ...

Right, there are many really good reasons we should be excited about DL-based models.
Before we delve into Deep Learning (DL) models ...

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

But, there are also many good reasons we want to know nice non-DL research performed before.
Before we delve into Deep Learning (DL) models ...

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

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

Also, it is always intriguing to think how the final NLI models (if any) would look like, or at least, what's the limitations of existing DL models.
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., cross-sentence attention). (*Full models*)
  - **Category II**: NLI models that do not use cross-sentence information. (*Sentence-vector-based models*)
  
  ■ This category of models 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 unsupervised pretraining
- Sentence-vector-based NLI models
  - A top-ranked model in RepEval-2017 Shared Task
  - Current top model based on dynamic self-attention
- Several additional topics
“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 unsupervised pretraining

Sentence-vector-based NLI models
  - A top-ranked model in RepEval-2017 Shared Task
  - Current top model based on dynamic self-attention

Several additional topics
Enhanced Sequential Inference Models (ESIM)

**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.

**Layer 2: Local Inference Modeling**
Collect information to perform “local” inference between words or phrases. (Some heuristics works well in this layer.)

**Layer 3: Inference Composition/Aggregation**
Perform composition/aggregation over local inference output to make the global judgement.
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 Modeling
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 $b$:

$$a = (a_1, \ldots, a_{\ell_a})$$
$$b = (b_1, \ldots, b_{\ell_b})$$

we can apply different encoders (e.g., here BiLSTM):

where $\bar{a}_i$ denotes the output vector of BiLSTM at the position $i$ of premise, which encodes word $a_i$ and its context.
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 Modeling
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 Modeling

Premise: Two dogs are running through a field

Hypothesis: There are animals outdoors
Local Inference Modeling

Premise

Two dogs are running through a field

Hypothesis

There are animals outdoors

Attention content

\[ \tilde{a}(\text{"dogs"}) = 0.05 \times \text{"There"} + 0.05 \times \text{"are"} + 0.8 \times \text{"animals"} + 0.1 \times \text{"outdoors"} \]

Attention Weights
Local Inference Modeling

Premise

Two dogs are running through a field

Hypothesis

There are animals outdoors

Attention content

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

Attention Weights
Local Inference Modeling

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

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

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

where,

\[
e_{ij} = \tilde{a}_i^T \tilde{b}_j
\]

(ESIM tried several more complicated functions of \(e_{ij} = f(\tilde{a}_i, \tilde{b}_j)\), which did not further help.)
Local Inference Modeling

- 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:
  \[
  m_a = [\tilde{a}; \tilde{a}; \tilde{a} - \tilde{a}; \tilde{a} \odot \tilde{a}]
  \]
  \[
  m_b = [\tilde{b}; \tilde{b}; \tilde{b} - \tilde{b}; \tilde{b} \odot \tilde{b}]
  \]
  - For premise, at each time step \(i\), concatenate \(\tilde{a}_i\) and \(\tilde{a}_i\), together with their:
    - element-wise product,
    - element-wise difference.

(The same is performed for the hypothesis.)
Some questions so far ...

• Some questions:
  ○ Instead of using chain RNN, how about other NN architectures?
  ○ How if one has access to more knowledge than that 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 Modeling
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:

\[ v_a = \text{BiLSTM}(m_a) \]
\[ v_b = \text{BiLSTM}(m_b) \]

where

\[ m_a = [a; \hat{a}; \bar{a} - \hat{a}; a \odot \hat{a}] \]
\[ m_b = [\tilde{b}; \hat{b}; \bar{b} - \hat{b}; b \odot \hat{b}] \]

● Then by concatenating the average and max-pooling of \( m_a \) and \( m_b \), we obtain a vector \( v \) which is fed to a classifier.
## Performance of ESIM on SNLI

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

Accuracy of ESIM and previous models on SNLI
Models Enhanced with Syntactic Structures
Models Enhanced with Syntactic Structures

- Syntax has been used in many non-neural NLI/RTE systems (MacCartney, ‘09; Dagan et al. ‘13).
- How to explore syntactic structures in NN-based NLI systems?

Several typical models:

- **Hierarchical Inference Models (HIM)** (Chen et al., ‘17) (full model)
- **Stack-augmented Parser-Interpreter Neural Network (SPINN)** (Bowman et al., ‘16) and follow-up work (sentence-vector-based models)
- **Tree-Based CNN (TBCNN)** (Mou et al., ‘16) (sentence-vector-based models)

MacCartney ‘09, Dagan et al. ‘13, Bowman et al. ‘16, Mou et al. ‘16, Chen et al. ‘17
Parse information can be considered in different phases of NLI.
Tree LSTM

Chain LSTM

Tree LSTM

Zhu et al. ‘15, Tai et al. ‘15, Le & Zuidema ‘15
Parse information can be first used to encode input sentences.
Attention weights showed that the tree models aligned “sitting down” with “standing” and the classifier relied on that to make the correct judgement.

The sequential model, however, soft-aligned “sitting” with both “reading” and “standing” and confused the classifier.
Perform “composition” on local inference information over trees:

\[
v_{a,t} = \text{TrLSTM}(F(m_{a,t}), h_{t-1}^L, h_{t-1}^R) \\
v_{b,t} = \text{TrLSTM}(F(m_{b,t}), h_{t-1}^L, h_{t-1}^R)
\]

where, \(m_{a,t}\) and \(m_{b,t}\) are first passed through a feed-forward layer \(F(\cdot)\) to reduce the number of parameters and alleviate overfitting.

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

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

**Ablation Analysis**

(The numbers are classification accuracy.)
Evans et al. (2018) constructed a dataset and explored deep learning models for detecting entailment in formal logic:

$p \models p \lor q \quad \neg p \land \neg q \models \neg q \quad p \not\models \neg q \quad \neg p \land \neg q \not\models p \lor q$

- 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 have been made to avoid annotation artifacts.
  - E.g. positive (entailment) and negative (non-entailment) examples must have the same distribution w.r.t. length of the formulae.
The results suggested that, if the structure of input is given, unambiguous, and a central feature of the task, models that explicitly exploit structures (e.g., treeLSTM) outperform models which must implicitly model the structure of sequences.
SPINN: Doing Away with Test-Time Tree

- Shift-reduce parser:
  - *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.**

During training time, trees offer supervision for shift-reduce parser. No need for test time trees!

*Image credit: Sam Bowman and co-authors.*
● Word vectors start on buffer.
● **Shift:** moves word vectors from buffer to stack.
● **Reduce:** pops top two vectors off the stack, applies $f^R: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, 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.
Do Latent Tree Learning Identify Meaningful Structure?

- Williams et al. (2018) conducted a comprehensive comparison on models that use explicit linguistic tree and latent trees.
  - The models include those proposed by Yogatama et al. (2017) and Choi et al. (2018) as well as variants of SPINN.
- Their main findings include:
  - “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, Choi et al. ‘18, Yogatama et al. ‘17
Intermission

Slides: nlitutorial.github.io
Models Enhanced with Semantic Roles
Recent research (Zhang et al., ‘19) incorporated Semantic Role Labeling (SRL) into NLI and found it improved the performance. The proposed model simply concatenated SRL embedding into word embedding.
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., ‘18), and BERT (Devlin et al., ‘18).

- 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. (*Feature-based pretrained models*)

- GPT and BERT: pretrained architectures and parameters are both used to perform NLI, parameters are finetuned in NLI, and otherwise no NLI-specific models/components are further used. (*Finetuning-based pretrained models*)

_Peters et al. ‘17, Radford et al. ‘18, Devlin et al. ‘18_
## Models Enhanced with Semantic Roles

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIIN</td>
<td>88.0</td>
</tr>
<tr>
<td>DR-BiLSTM</td>
<td>88.5</td>
</tr>
<tr>
<td>CAFE</td>
<td>88.5</td>
</tr>
<tr>
<td>MAN</td>
<td>88.3</td>
</tr>
<tr>
<td>KIM</td>
<td>88.6</td>
</tr>
<tr>
<td>DMAN</td>
<td>88.8</td>
</tr>
<tr>
<td>ESIM + TreeLSTM</td>
<td>88.6</td>
</tr>
<tr>
<td>ESIM + ELMo</td>
<td>88.7</td>
</tr>
<tr>
<td>DCRCN</td>
<td>88.9</td>
</tr>
<tr>
<td>LM-Transformer</td>
<td>89.9</td>
</tr>
<tr>
<td>MT-DNN†</td>
<td>91.1</td>
</tr>
<tr>
<td>Baseline (ELMo) + SRL</td>
<td>89.1</td>
</tr>
<tr>
<td>Baseline (BERT\textsubscript{BASE}) + SRL</td>
<td>89.6</td>
</tr>
<tr>
<td>Baseline (BERT\textsubscript{LARGE}) + SRL</td>
<td>90.4</td>
</tr>
<tr>
<td>+ SRL</td>
<td>91.3</td>
</tr>
</tbody>
</table>

**Accuracy on SNLI**

*Zhang et al. ‘19*
Modeling External Knowledge

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

● leveraging structured (often human-curated) knowledge

● using unsupervisedly pretrained models
Modeling External Knowledge

Leveraging Structured Knowledge
NLI Models Enhanced with External Knowledge: The KIM Model

Overall architecture of Knowledge-based Inference Model (KIM) (Chen et al. ‘18)
NLI Models Enhanced with External Knowledge: The KIM Model

- Knowledge-enhanced co-attention:

\[ e_{ij} = (a_i^s)^T b_j^s + F(r_{ij}) \]

- Intuitively lexical semantics such as synonymy, antonymy, hypernymy, and co-hyponymy may help soft-align a premise to its hypothesis.

- Specifically, \( r_{ij} \) is a vector of semantic relations between \( i^{th} \) word in a premise and \( j^{th} \) word in its hypothesis. The relations can be extracted from resources such as WordNet/ConceptNet or embedding learned from a knowledge graph.

Chen et al. '18
Local inference with external knowledge:

\[ a_i^m = G([a_i^s; a_i^c; a_i^s - a_i^c; a_i^s \odot a_i^c; \sum_{j=1}^{\ell_b} \alpha_{ij} r_{ij}]) , \]

- In addition to helping soft-alignment, external knowledge can also bring richer entailment information that does not exist in training data.

Enhancing inference composition/aggregation:

\[ a^w = \sum_{i=1}^{\ell_a} \frac{\exp(H(\sum_{j=1}^{\ell_b} \alpha_{ij} r_{ij})) a_i^v}{\sum_{i=1}^{\ell_a} \exp(H(\sum_{j=1}^{\ell_b} \alpha_{ij} r_{ij}))} , \]
## Accuracy on SNLI

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

Performance of KIM under different sizes of training-data.

Performance of KIM under different amounts of external knowledge.

Chen et al. ‘18
For a premise in SNLI, Glockner et al. (2018) generated a hypothesis by replacing a single word in the premise.

The aim is to help test if a NLI systems can actually learn simple lexical and word knowledge.

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

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

KIM performs much better than other models on this dataset.

---

**Accuracy on the Glockner Dataset**

<table>
<thead>
<tr>
<th>Model</th>
<th>SNLI</th>
<th>Glockner’s (Δ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Parikh et al., 2016)*</td>
<td>84.7</td>
<td>51.9 (-32.8)</td>
</tr>
<tr>
<td>(Nie and Bansal, 2017)*</td>
<td>86.0</td>
<td>62.2 (-23.8)</td>
</tr>
<tr>
<td>ESIM *</td>
<td>87.9</td>
<td>65.6 (-22.3)</td>
</tr>
<tr>
<td>KIM (This paper)</td>
<td>88.6</td>
<td>83.5 (-5.1)</td>
</tr>
</tbody>
</table>

---

Glockner et al. ‘18
Modeling External Knowledge

Leveraging Unsupervised Pretraining
Pretrained models can leverage large unannotated datasets, which have brought forward the state of the art of NLI and many other tasks.

- See (Peters et al., ‘17, Radford et al., ‘18, Devlin et al., ‘18) for more details.

- Whether/how the models using human-curated structured knowledge (e.g., KIM) and those using unsupervised pretraining (e.g., BERT) complement each other?
External Knowledge: BERT vs. KIM

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 in which KIM is correct but BERT is wrong.

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 in which BERT is correct but KIM is wrong.
More Analysis on Pairs of Systems

<table>
<thead>
<tr>
<th></th>
<th>BERT</th>
<th>GPT</th>
<th>KIM</th>
<th>ESIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>.561</td>
<td>.580</td>
<td>.652</td>
<td>.616</td>
</tr>
<tr>
<td>GPT</td>
<td>.304</td>
<td>.543</td>
<td>.457</td>
<td></td>
</tr>
<tr>
<td>KIM</td>
<td>.491</td>
<td></td>
<td>.552</td>
<td></td>
</tr>
<tr>
<td>ESIM</td>
<td></td>
<td></td>
<td>.320</td>
<td></td>
</tr>
</tbody>
</table>

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).

- BERT and KIM seem to complement each other more than other pairs, e.g., BERT and GPT.
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 representation quality for individual sentence.
The RepEval-2017 Shared Task (Nangia et al., ‘17) adopted the MNLI dataset to evaluate sentence representation.

- We will discuss one of the top-ranked models (Chen et al., ‘17b). Other top models can be found in (Nie and Bansal, ‘17; Balazs et al., ‘17).
RNN-Based Inference Model with Gated Attention

Chen et al. '17b
In addition to average 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 computed using the input, forget, and output gates of the top-layer BiLSTM.
Results

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

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., ‘18b) for more work on generalized pooling.

Chen et al. ‘18b
CNN with Dynamic Self-Attention

- So far, the model proposed by Yoon et al. (2018) achieves the best performance on SNLI among sentence-vector-based models.
- Key idea: stacks a dynamic self-attention over CNN (with dense connection)
- The proposed dynamic self-attention borrows ideas from the Capsule Network (Sabour et al. ‘17; Hinton et al., ‘18).

Yoon et al. ‘18, Sabour et al. ‘17, Hinton et al. ‘18
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., the face’s 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

- Key differences:
  - Input of a capsule cell is a number of vectors ($u_1$ is a vector) but not a scalar ($x_1$ is a scalar).
  - Voting parameters $c_1$, $c_2$, $c_3$ are not part of model parameters — they are 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($\hat{u}_{j|i}$, $r$, $l$)
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 $l$: $c_i \leftarrow \text{softmax}(b_i)$
5: for all capsule $j$ in layer $(l + 1)$: $s_j \leftarrow \sum_i c_{ij} \hat{u}_{j|i}$
6: for all capsule $j$ in layer $(l + 1)$: $v_j \leftarrow \text{squash}(s_j)$
7: for all capsule $i$ in layer $l$ and capsule $j$ in layer $(l + 1)$: $b_{ij} \leftarrow b_{ij} + \hat{u}_{j|i} \cdot v_j$

return $v_j$

- 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
The proposed model borrows the idea of weight adaptation method in dynamic routing to adapt attention weight $a_{ij}$. (Note that in dynamic self-attention, weights are normalized along lower-level vectors, indexed by $k$, while in dynamic routing in CapsuleNet normalization is performed along higher-level vectors/capsules.)

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

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

Current leaderboard of sentence-vector-based models on SNLI (as of June 1\textsuperscript{st}, 2019).
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
Revisiting Artifacts of Data
As discussed above, Glockner et al. (2018) create a new test set that shows the deficiency of NLI systems in modeling 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.

<table>
<thead>
<tr>
<th>Premise/Hypothesis</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>The man is holding a saxophone</td>
<td>contradiction</td>
</tr>
<tr>
<td>The man is holding an electric guitar</td>
<td></td>
</tr>
<tr>
<td>A little girl is very sad.</td>
<td>entailment</td>
</tr>
<tr>
<td>A little girl is very unhappy.</td>
<td></td>
</tr>
<tr>
<td>A couple drinking wine</td>
<td>neutral</td>
</tr>
<tr>
<td>A couple drinking champagne</td>
<td></td>
</tr>
</tbody>
</table>

Glockner et al. ‘18
The performance of NLI systems on the new test set is substantially worse, suggesting some drawback of the existing NLI systems/datasets in actually modelling NLI.

Accuracy of models on SNLI and the Glockner dataset.

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

- Naik et al. (2018) proposed 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, and word overlap.
  - Noise test: robustness on “spelling errors”.

Naik et al. ‘18
### “Stress Tests” for NLI

<table>
<thead>
<tr>
<th>System</th>
<th>Original MultiNLI Dev</th>
<th>Competence Test</th>
<th>Distraction Test</th>
<th>Noise Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mat  Mis</td>
<td>Mat  Mis</td>
<td>Mat  Mis</td>
<td>Mat  Mis</td>
</tr>
<tr>
<td>NB</td>
<td>74.2 74.8</td>
<td>15.1 19.3</td>
<td>47.2 47.1</td>
<td>51.1 49.8</td>
</tr>
<tr>
<td>CH</td>
<td>73.7 72.8</td>
<td>11.6 9.3</td>
<td>58.3 58.4</td>
<td>68.3 69.1</td>
</tr>
<tr>
<td>RC</td>
<td>71.3 71.6</td>
<td>36.4 32.8</td>
<td>53.7 54.4</td>
<td>66.6 67.0</td>
</tr>
<tr>
<td>IS</td>
<td>70.3 70.6</td>
<td>14.4 10.2</td>
<td>50.0 50.2</td>
<td>58.3 59.4</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>70.2 70.8</td>
<td>13.2 9.8</td>
<td>57.0 58.5</td>
<td>65.0 65.1</td>
</tr>
<tr>
<td>CBOV</td>
<td>63.5 64.2</td>
<td>6.3 3.6</td>
<td>53.6 55.6</td>
<td>60.3 60.6</td>
</tr>
</tbody>
</table>

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., ‘17b), and **RC** (Balazs et al., ‘17) are models submitted to RepEvel-2017. **IS** (Conneau et al., ‘17) is a model proposed to learn general sentence embedding trained on NLI.
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.
- Models should have comparable accuracy on the original test set and swapped test set for contradiction and neutral.
Swapping Premise and Hypothesis

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

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 datasets can be found in Poliak et. al, ‘18, Talman and Chatzikyriakidis, ‘19.
Bringing Explanation to NLI
e-SNLI: Bringing Explanation to NLI

- e-SNLI extends SNLI with an additional layer of human-annotated natural language explanation.
- 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 datasets.

Camburu et al. ‘18
e-SNLI: Bringing Explanation to NLI

- **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.

Overview of the e-INFERSENT architecture.
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.
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”
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
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)*
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
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)
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:

Performance Estimates (Several Sources!)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Most Frequent Class</th>
<th>SotA</th>
<th>Human (approx.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNLI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SciTail</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MultiNLI</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend:
- Most Frequent Class
- SotA
- Human (approx.)
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 construction processing used to construct the MNLI corpora.

- Can be used to evaluate both cross-lingual NLI models and multilingual text embedding models.

Conneau et al. ‘18
XNLI: Evaluating Cross-lingual Sentence Representations

<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>fr</th>
<th>es</th>
<th>de</th>
<th>el</th>
<th>bg</th>
<th>ru</th>
<th>tr</th>
<th>ar</th>
<th>vi</th>
<th>th</th>
<th>zh</th>
<th>hi</th>
<th>sw</th>
<th>ur</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Machine translation baselines (TRANSLATE TRAIN)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTM-last</td>
<td>71.0</td>
<td>66.7</td>
<td>67.0</td>
<td>65.7</td>
<td>65.3</td>
<td>65.6</td>
<td>61.9</td>
<td>63.9</td>
<td>63.1</td>
<td>61.3</td>
<td>65.7</td>
<td>61.3</td>
<td>55.2</td>
<td>55.2</td>
<td></td>
</tr>
<tr>
<td>BiLSTM-max</td>
<td><strong>73.7</strong></td>
<td>68.3</td>
<td>68.8</td>
<td>66.5</td>
<td>66.4</td>
<td>67.4</td>
<td>66.5</td>
<td>64.5</td>
<td>65.8</td>
<td>66.0</td>
<td>62.8</td>
<td>67.0</td>
<td>62.1</td>
<td>58.2</td>
<td>56.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>fr</th>
<th>es</th>
<th>de</th>
<th>el</th>
<th>bg</th>
<th>ru</th>
<th>tr</th>
<th>ar</th>
<th>vi</th>
<th>th</th>
<th>zh</th>
<th>hi</th>
<th>sw</th>
<th>ur</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Machine translation baselines (TRANSLATE TEST)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTM-last</td>
<td>71.0</td>
<td>68.3</td>
<td>68.7</td>
<td>66.9</td>
<td>67.3</td>
<td>68.1</td>
<td>66.2</td>
<td>64.9</td>
<td>65.8</td>
<td>64.3</td>
<td>63.2</td>
<td>66.5</td>
<td>61.8</td>
<td>60.1</td>
<td>58.1</td>
</tr>
<tr>
<td>BiLSTM-max</td>
<td><strong>73.7</strong></td>
<td>70.4</td>
<td>70.7</td>
<td>68.7</td>
<td>69.1</td>
<td>70.4</td>
<td>67.8</td>
<td>66.3</td>
<td>66.8</td>
<td>66.5</td>
<td>64.4</td>
<td>68.3</td>
<td>64.2</td>
<td>61.8</td>
<td>59.3</td>
</tr>
</tbody>
</table>

**Evaluation of XNLI multilingual sentence encoders (in-domain)**

<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>fr</th>
<th>es</th>
<th>de</th>
<th>el</th>
<th>bg</th>
<th>ru</th>
<th>tr</th>
<th>ar</th>
<th>vi</th>
<th>th</th>
<th>zh</th>
<th>hi</th>
<th>sw</th>
<th>ur</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-BiLSTM-last</td>
<td>71.0</td>
<td>65.2</td>
<td>67.8</td>
<td>66.6</td>
<td>66.3</td>
<td>65.7</td>
<td>63.7</td>
<td>64.2</td>
<td>62.7</td>
<td>65.6</td>
<td>62.7</td>
<td>63.7</td>
<td>62.8</td>
<td>54.1</td>
<td>56.4</td>
</tr>
<tr>
<td>X-BiLSTM-max</td>
<td><strong>73.7</strong></td>
<td>67.7</td>
<td>68.7</td>
<td>67.7</td>
<td>68.9</td>
<td>67.9</td>
<td>65.4</td>
<td>64.2</td>
<td>64.8</td>
<td>66.4</td>
<td>64.1</td>
<td>65.8</td>
<td>64.1</td>
<td>55.7</td>
<td>58.4</td>
</tr>
</tbody>
</table>

**Evaluation of pretrained multilingual sentence encoders (transfer learning)**

<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>fr</th>
<th>es</th>
<th>de</th>
<th>el</th>
<th>bg</th>
<th>ru</th>
<th>tr</th>
<th>ar</th>
<th>vi</th>
<th>th</th>
<th>zh</th>
<th>hi</th>
<th>sw</th>
<th>ur</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-CBOW</td>
<td>64.5</td>
<td>60.3</td>
<td>60.7</td>
<td>61.0</td>
<td>60.5</td>
<td>60.4</td>
<td>57.8</td>
<td>58.7</td>
<td>57.5</td>
<td>58.8</td>
<td>56.9</td>
<td>58.8</td>
<td>56.3</td>
<td>50.4</td>
<td>52.2</td>
</tr>
</tbody>
</table>

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

Conneau et al. ‘18, Lample & Conneau. ‘19
The Discourse Marker Augmented Network (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 Discourse Marker Prediction (Nie et al., 2017) is incorporated into DMAN through a reinforcement learning component.