Modeling Natural Language Semantics with Learned Representations

Samuel R. Bowman

Premise: A man speaking or singing into a microphone while playing the piano.
Hypothesis: A man is performing surgery on a giraffe while singing.
Label: contradiction
Goal: Build computational models that can learn to understand and reason with human language.
Open problems in NLP

Question answering

How old is the oldest leader of an OPEC country?

+ [Image] = ?

Summarization

= Drug X interacts badly with drug Y.
Open problems in linguistics

What prior knowledge must a learner have in order to fully learn language?
Open problems at the intersection

How do we combine logical approaches to meaning with a rich representations of word meaning?

∀x. ∃y...

If all dogs bark, do most puppies make sounds?

Is a labrador more of a dog than a chihuahua?
Neural networks in NLP

- **2010:** Marginal
- **2016:** Major research area
  
  Standard for parsing, classification, ...

Figure from Christoph Burgmer.
Neural machine translation

Sutskever et al. ‘14, Bahdanau et al. ‘15, Luong et al. ‘15 (figure from Chris Manning)
Today: Some open questions

Goal: Build neural network models that can learn to understand and reason with human language.

- Can continuous models do symbolic reasoning?
- Can they learn to understand real language?
- What can formal semantics teach them?
Background:
Neural networks and natural language
Distributed feature vectors for words

\[
good \quad \mapsto \quad < 0.9, -0.2 > \\
okay \quad \mapsto \quad < 0.8, -0.5 > \\
bad \quad \mapsto \quad < 0.2, -0.7 > \\
\ldots
\]
Composition: From words to sentences

How do we construct sentence representations from word representations?
Composition: From words to sentences

Sequence-based (recurrent) neural network encoder

Rumelhart et al., ‘86; Werbos, ‘90; Mikolov, ‘10
Composition: From words to sentences

Alternative model: *Tree-structured (recursive) neural network encoder*

Goller & Küchler '96; Socher et al. '10
Some open questions

Goal: Build neural network models that can learn to understand and reason with human language.

● Can continuous models do symbolic reasoning?
● Can they learn to understand real language?
● What can formal semantics teach them?
Measuring success

Goal: Build neural network models that can learn to understand and reason with human language.

What does success look like?

Where does supervision come from?
Natural language inference (NLI)
or recognizing textual entailment (RTE)

James Byron Dean refused to move without blue jeans
{entails, contradicts, neither}

James Dean didn’t dance without pants
Natural language inference (NLI)

or recognizing textual entailment (RTE)

James Byron Dean refused to move without blue jeans

{entails, contradicts, neither}

James Dean didn’t dance without pants
Why natural language inference?

James Byron Dean refused to move without blue jeans
\{entails, contradicts, neither\}

James Dean didn’t dance without pants

- move vs. dance (hyponym and hyponymy)
- refused to vs. didn’t (factives and implicatives)
- James B. Dean vs. James Dean (coreference)

...
Why natural language inference?

Natural language inference is a major sub-problem of:

- Question answering
- Semantic web search
- Summarization
- Machine translation
  and more!
NLI and Natural Logic

Research in **Natural Logic** formally characterizes sound inference patterns over natural language.

\[
dance \sqsubseteq move
\]

so...

James Dean danced \sqsubseteq James Dean moved

but...

James Dean \textit{didn't} dance \sqsupseteq James Dean \textit{didn't} move

Sánchez-Valencia, ‘91; MacCartney, ‘09; Icard & Moss ‘13
Reasoning with words
Building a learning problem

Training data

dance entails move
tango entails dance
sleep contradicts dance
waltz entails dance

Test data

sleep ? waltz
Natural logic: The seven relations

Seven possible relations between phrases/sentences:

<table>
<thead>
<tr>
<th>Relation</th>
<th>Example 1</th>
<th>Symbol</th>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>equivalence</td>
<td>couch</td>
<td>≡</td>
<td>sofa</td>
</tr>
<tr>
<td>forward entailment (strict)</td>
<td>crow</td>
<td>⊏</td>
<td>bird</td>
</tr>
<tr>
<td>reverse entailment (strict)</td>
<td>European</td>
<td>⊐</td>
<td>French</td>
</tr>
<tr>
<td>negation (exhaustive exclusion)</td>
<td>human</td>
<td>^</td>
<td>nonhuman</td>
</tr>
<tr>
<td>alternation (non-exhaustive exclusion)</td>
<td>cat</td>
<td></td>
<td>dog</td>
</tr>
<tr>
<td>cover (exhaustive non-exclusion)</td>
<td>animal</td>
<td>⌂</td>
<td>nonhuman</td>
</tr>
<tr>
<td>independence</td>
<td>hungry</td>
<td>#</td>
<td>hippo</td>
</tr>
</tbody>
</table>
Lexical relation data

<table>
<thead>
<tr>
<th>TRAIN</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a \equiv a$</td>
<td>$a \equiv b$</td>
</tr>
<tr>
<td>$a \wedge f$</td>
<td>$a \dashv d$</td>
</tr>
<tr>
<td>$b \dashv c$</td>
<td>$a \sqsubseteq e$</td>
</tr>
<tr>
<td>$b \dashv d$</td>
<td>$b \sqsubseteq e$</td>
</tr>
</tbody>
</table>
The simplest viable neural inference model

\[ P(\mathcal{E}) = 0.8 \]
Learning lexical relations

Generalization (test) accuracy 99.6%

Training

dance *entails* move

Test

sleep ? waltz
Reasoning with novel sentences
## Function words and infinite languages

<table>
<thead>
<tr>
<th>TRAIN</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>b   $\equiv$   b</td>
<td>not a $\wedge$  a</td>
</tr>
<tr>
<td>not (not a) $\equiv$ a</td>
<td>c or d $\sqcap$ d</td>
</tr>
<tr>
<td>c $\square$ b and c</td>
<td>not not b $\equiv$ b</td>
</tr>
<tr>
<td></td>
<td>not (not a and not d) $\equiv$ a or d</td>
</tr>
</tbody>
</table>
The model: A TreeRNN for NLI

\[ P(\Downarrow) = 0.8 \]
Function words and infinite languages

Accuracy

Size of longer expression

Most frequent class
An example with twelve connectives

\[
((\neg d) \lor (\neg ((\neg (b \lor e)) \land (b \lor (\neg b))))))
\]

\[
(\neg ((\neg ((b \land (\neg b)) \lor (\neg (d \land b)))) \lor (\neg (((\neg e) \lor d) \land (d \lor c))))))
\]
Function words and infinite languages

![Graph showing accuracy of longest expression sizes]
Function words and infinite languages

![Graph showing the accuracy of function words and infinite languages over the size of longer expression. The graph compares the most frequent class and 50d TreeRNN accuracy.]
Some open questions

Goal: Build neural network models that can learn to understand and reason with human language.

- Can continuous models do symbolic reasoning?
- Can they learn to understand real language?
- What can formal semantics teach them?
## What data can we learn from?

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Complete Sentences</th>
<th>Human Labeled</th>
<th>Size (num. pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FraCaS</td>
<td>✓</td>
<td>✓</td>
<td>.3k</td>
</tr>
<tr>
<td>RTE 1-5</td>
<td>✓</td>
<td>✓</td>
<td>7k</td>
</tr>
<tr>
<td>SICK</td>
<td>✓</td>
<td>✓</td>
<td>10k</td>
</tr>
<tr>
<td>DenotationGraph</td>
<td>✗</td>
<td>✗</td>
<td>728k</td>
</tr>
<tr>
<td>Levy Graphs</td>
<td>✗</td>
<td>✗</td>
<td>1,500k</td>
</tr>
<tr>
<td>PPDB 2.0</td>
<td>✗</td>
<td>✗</td>
<td>100,000k</td>
</tr>
<tr>
<td>Corpus</td>
<td>Complete Sentences</td>
<td>Human Labeled</td>
<td>Size (num. pairs)</td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------------</td>
<td>---------------</td>
<td>------------------</td>
</tr>
<tr>
<td>FraCaS</td>
<td>✓</td>
<td>✓</td>
<td>.3k</td>
</tr>
<tr>
<td>RTE 1-5</td>
<td>✓</td>
<td>✓</td>
<td>7k</td>
</tr>
<tr>
<td>SICK</td>
<td>✓</td>
<td>✓</td>
<td>10k</td>
</tr>
<tr>
<td>DenotationGraph</td>
<td>✗</td>
<td>✗</td>
<td>728k</td>
</tr>
<tr>
<td>Levy Graphs</td>
<td>✗</td>
<td>✗</td>
<td>1,500k</td>
</tr>
<tr>
<td>PPDB 2.0</td>
<td>✗</td>
<td>✗</td>
<td>100,000k</td>
</tr>
</tbody>
</table>
Training neural networks on existing data

A little girl is looking at a woman in costume
{entailment, contradiction, neutral}
The little girl is looking at a man in costume

Approach                      SICK test acc.
Just guessing ‘neutral’        56.7%
Best NN model                  76.9%
Best prior non-NN model        84.5%
## What data can we learn from?

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Complete Sentences</th>
<th>Human Labeled</th>
<th>Size (num. pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FraCaS</td>
<td>✓</td>
<td>✓</td>
<td>.3k</td>
</tr>
<tr>
<td>RTE 1-5</td>
<td>✓</td>
<td>✓</td>
<td>7k</td>
</tr>
<tr>
<td>SICK</td>
<td>✓</td>
<td>✓</td>
<td>10k</td>
</tr>
<tr>
<td>DenotationGraph</td>
<td>✗</td>
<td>✗</td>
<td>728k</td>
</tr>
<tr>
<td>Levy Graphs</td>
<td>✗</td>
<td>✗</td>
<td>1,500k</td>
</tr>
<tr>
<td>PPDB 2.0</td>
<td>✗</td>
<td>✗</td>
<td>100,000k</td>
</tr>
</tbody>
</table>
## Our large, human-labeled NLI corpus

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Complete Sentences</th>
<th>Human Labeled</th>
<th>Size (num. pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FraCaS</td>
<td>✓</td>
<td>✓</td>
<td>.3k</td>
</tr>
<tr>
<td>RTE 1-5</td>
<td>✓</td>
<td>✓</td>
<td>7k</td>
</tr>
<tr>
<td>SICK</td>
<td>✓</td>
<td>✓</td>
<td>10k</td>
</tr>
<tr>
<td>SNLI</td>
<td>✓</td>
<td>✓</td>
<td>570k</td>
</tr>
<tr>
<td>DenotationGraph</td>
<td>✗</td>
<td>✗</td>
<td>728k</td>
</tr>
<tr>
<td>Levy Graphs</td>
<td>✗</td>
<td>✗</td>
<td>1,500k</td>
</tr>
<tr>
<td>PPDB 2.0</td>
<td>✗</td>
<td>✗</td>
<td>100,000k</td>
</tr>
</tbody>
</table>

Bowman et al. ‘15: “A large annotated corpus for learning natural language inference”
Girl in a red coat, blue head wrap and jeans is making a snow angel.

\{entailment, contradiction, neutral\}

A girl outside plays in the snow.

- Typical examples require:
  - Full sentence understanding.
  - Common sense world knowledge.
- Outside the scope of pure natural logic.
How do we collect this data?

Prompt for Mechanical Turk annotators:

We will show you the caption for a photo. We will not show you the photo. Using just the caption and what you know about the world, write a new caption for the same photo that is {definitely accurate, definitely inaccurate, possibly accurate}. 
## Initial machine learning results

<table>
<thead>
<tr>
<th>Model</th>
<th>Test acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Just guessing ‘entailment’</td>
<td>33.7%</td>
</tr>
<tr>
<td>Big simple classifier</td>
<td>78.2%</td>
</tr>
<tr>
<td>Recurrent (sequence) NN model</td>
<td>77.6%</td>
</tr>
</tbody>
</table>
Extramural results

- **Sep. 2015**: Corpus release
- **Sep. 2015**: Google DeepMind/UCL/Oxford
- **Nov. 2015**: U. of Toronto
- **Dec. 2015**: Peking U./Baidu
- **Dec. 2015**: Singapore Management U.
- **Jan. 2016**: U. of Edinburgh
- **Feb. 2016**: Unbabel Lda./IT/INESC-ID (Pt.)
Some open questions

Goal: Build neural network models that can learn to understand and reason with human language.

- Can continuous models do symbolic reasoning?
- Can they learn to understand real language?
- What can formal semantics teach them?
Working assumptions in formal semantics

Loosely, the principle of compositionality:

the    cat    sat    down

Sentence meanings are constructed incrementally by composing together word meanings.
Working assumptions in formal semantics

Loosely, *the principle of compositionality*:

This composition process can be most concisely described using a phrase structure that *roughly* follows the phrase structure used in syntax.
Recursion with propositional logic

Accuracy

Size of longer expression

Most frequent class

50d TreeRNN
Recursion with propositional logic

![Graph showing the accuracy of different models as a function of the size of a longer expression. The x-axis represents the size of the longer expression, and the y-axis represents accuracy on a percentage scale. Three lines are plotted: one for the most frequent class, one for 50d TreeRNN, and one for 50d LSTM RNN. The graph illustrates how the accuracy decreases as the size of the expression increases.]
Tree structured models in practice

Robust successes on NLP for tasks with smaller datasets: sentiment analysis, paraphrase detection...

Larger datasets? Too slow.
Batched computation
Batched computation

the perfect dessert was red curry dessert was not that great
Transition-based parsing

Is it possible to do tree-structured compositionality in an efficient model?

Transition-based parsing offers a clue.

Bowman et al. ‘16: “A Fast Unified Model for Parsing and Sentence Understanding”
Transition-based parsing

SHIFT SHIFT
REDUCE SHIFT
SHIFT REDUCE
REDUCE

SHIFT SHIFT
SHIFT SHIFT
REDUCE REDUCE
REDUCE

SHIFT SHIFT
SHIFT REDUCE
SHIFT REDUCE
REDUCE
Transition-based parsing
Transition-based parsing

Stack

Buffer

SHIFT

down

REDUCE

the cat

sat

down

REDUCE

the cat

sat down

REDUCE

(the cat) (sat down)
Stack-augmented Parser-Interpreter NN
### The shift-reduce model on SNLI

<table>
<thead>
<tr>
<th>Model</th>
<th>Test acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence model (our prev. impl.):</td>
<td>77.6%</td>
</tr>
<tr>
<td>Best comparable model:</td>
<td>82.1%</td>
</tr>
<tr>
<td>Sequence model (our new impl.):</td>
<td>80.6%</td>
</tr>
<tr>
<td>SPINN (purely tree-structured):</td>
<td>80.9%</td>
</tr>
<tr>
<td>SPINN (hybrid):</td>
<td><strong>83.2%</strong></td>
</tr>
</tbody>
</table>
Ongoing work: Future directions

Neural attention
State-of-the-art attention-based model: 89.0%
Attention-based SPINN: ?

Learning syntax from semantics
Build models that can learn to use whatever parse structure best supports the task at hand
Some open questions, and some answers

Goal: Build neural network models that can learn to understand and reason with human language.

Can continuous models do symbolic reasoning?
- Yes, e.g., lexical relations, recursive functions...

\[
\begin{align*}
&((\neg d) \lor (\neg((\neg(b \lor e)) \land (b \lor (\neg b)))))) \\
\quad &\quad \therefore \\
&\neg((\neg((b \land (\neg b)) \lor (\neg(d \land b)))) \lor (\neg(((\neg(e) \lor d) \land (d \lor c))))))
\end{align*}
\]
Some open questions, and some answers

Goal: Build neural network models that can learn to understand and reason with human language.

Can they learn to understand real language?
● Not perfectly yet, but at the state of the art and making rapid progress.

Girl in a red coat, blue head wrap and jeans is making a snow angel.

{entailment, contradiction, neutral}

A girl outside plays in the snow.
Some open questions, and some answers

Goal: Build neural network models that can learn to understand and reason with human language.

What can formal semantics teach them?

- Compositionality, at least: yields huge gains on artificial data, and significant gains on English.
Where we are now

Neural networks are the most effective tool we have for learning to understand natural language, but our models are still far from human-level understanding.
Future work

To fill the gap, more work is needed into:

● Discovering what aspects of meaning these models learn to use in practice.
● Applying our theoretical understanding of language to build helpful learning biases.
● Building models that can learn to refine their representations of meaning using raw text or other kinds of labeled data.
Modeling Natural Language Semantics with Learned Representations

Samuel R. Bowman

Premise: A man speaking or singing into a microphone while playing the piano.
Hypothesis: A man is performing surgery on a giraffe while singing.
Label: contradiction