Fast Breaking and Slow Building
of textual inference models

Vered Shwartz

December 2019
State-of-the-art AI solutions: (1) Google BERT, an AI model that understands language better than humans

Alibaba AI Beats Humans in Reading-Comprehension...
Alibaba Group’s machine-learning technology is better at reading comprehension than humans, according to a well-known test built for the industry by Microsoft.

Algorithms Have Nearly Mastered Human Language. Why Can’t They Stop Being Sexist?
It turns out that data-fueled algorithms are no better than humans—and ... Even AI researchers who work with machine learning models—like neural nets, which ...

Designed by freepik.com
What’s in this talk?

Breaking

- [Glockner et al., 2018]
- [Rozen et al., 2019]

Building

- [Shwartz et al., 2017]
- [Barhom et al., 2019]
Breaking

NLI

[Barhom et al., 2019]

[Shwartz et al., 2017]

[Rozen et al., 2019]

Coreference

Building
Breaking NLI Systems
with Sentences that Require Simple Lexical Inferences

Max Glockner$^1$, Vered Shwartz$^2$ and Yoav Goldberg$^2$

$^1$TU Darmstadt

$^2$Bar-Ilan University

ACL 2018
SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])
SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])
- Premises are image captions, hypotheses generated by crowdsourcing workers:

**Premise**

Street performer is doing his act for kids

**Hypotheses**

1. A person performing for children on the street ⇒ ENTAILMENT
2. A juggler entertaining a group of children on the street ⇒ NEUTRAL
3. A magician performing for an audience in a nightclub ⇒ CONTRADICTION
SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])
- Premises are image captions, hypotheses generated by crowdsourcing workers:

<table>
<thead>
<tr>
<th>Premise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Street performer is doing his act for kids</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hypotheses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A person performing for children on the street</td>
</tr>
</tbody>
</table>
SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])
- Premises are image captions, hypotheses generated by crowdsourcing workers:

**Premise**
Street performer is doing his act for kids

**Hypotheses**
1. A person performing for children on the street ⇒ **ENTAILMENT**
SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])
- Premises are image captions, hypotheses generated by crowdsourcing workers:

**Premise**
Street performer is doing his act for kids

**Hypotheses**
1. A person performing for children on the street ⇒ ENTAILMENT
2. A juggler entertaining a group of children on the street
SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])
- Premises are image captions, hypotheses generated by crowdsourcing workers:

<table>
<thead>
<tr>
<th>Premise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Street performer is doing his act for kids</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hypotheses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A person performing for children on the street ⇒ ENTAILMENT</td>
</tr>
<tr>
<td>2. A juggler entertaining a group of children on the street ⇒ NEUTRAL</td>
</tr>
</tbody>
</table>
SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])
- Premises are image captions, hypotheses generated by crowdsourcing workers:

**Premise**
Street performer is doing his act for kids

**Hypotheses**
1. A person performing for children on the street $\Rightarrow$ **ENTAILMENT**
2. A juggler entertaining a group of children on the street $\Rightarrow$ **NEUTRAL**
3. A magician performing for an audience in a nightclub
SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])
- Premises are image captions, hypotheses generated by crowdsourcing workers:

**Premise**
Street performer is doing his act for kids

**Hypotheses**
1. A person performing for children on the street ⇒ **ENTAILMENT**
2. A juggler entertaining a group of children on the street ⇒ **NEUTRAL**
3. A magician performing for an audience in a nightclub ⇒ **CONTRADICTION**

- Event co-reference assumption
Neural NLI Models

- End-to-end, either **sentence-encoding** or **attention-based**

![Diagram of Neural NLI Models]

- Label
- Classifier
- Extract Features
- Premise Encoder
- Hypothesis Encoder
- Premise
- Hypothesis

Lexical knowledge: only from pre-trained word embeddings
As opposed to using resources like WordNet
SOTA exceeds human performance...
Neural NLI Models

- End-to-end, either sentence-encoding or attention-based

![Diagram of Neural NLI Models](image)
Neural NLI Models

- End-to-end, either **sentence-encoding** or **attention-based**

- Lexical knowledge: only from pre-trained word embeddings
  - As opposed to using resources like WordNet
Neural NLI Models

- End-to-end, either **sentence-encoding** or **attention-based**

- Lexical knowledge: only from pre-trained word embeddings
  - As opposed to using resources like WordNet
- SOTA exceeds human performance...
Neural NLI Models

- End-to-end, either **sentence-encoding** or **attention-based**

- **Lexical knowledge:** only from pre-trained word embeddings
  - As opposed to using resources like WordNet
- **SOTA exceeds human performance...**

---

1
Neural NLI Models

End-to-end, either **sentence-encoding** or **attention-based**

Lexical knowledge: only from pre-trained word embeddings
  - As opposed to using resources like WordNet

SOTA exceeds human performance... ¹

¹[Gururangan et al., 2018, Poliak et al., 2018]: by learning “easy clues”
Annotation Artifacts in SNLI

- [Gururangan et al., 2018, Poliak et al., 2018]: good performance on SNLI based on the hypothesis alone
Annotation Artifacts in SNLI

- [Gururangan et al., 2018, Poliak et al., 2018]: good performance on SNLI based on the hypothesis alone
- This is a result of the annotation procedure
[Gururangan et al., 2018, Poliak et al., 2018]: good performance on SNLI based on the hypothesis alone. This is a result of the annotation procedure:

- Negation (not, never, nobody) is correlated with contradiction.
Annotation Artifacts in SNLI

- [Gururangan et al., 2018, Poliak et al., 2018]: good performance on SNLI based on the hypothesis alone
- This is a result of the annotation procedure
  - Negation (not, never, nobody) is correlated with *contradiction*
  - ...and “cat” as well (many dog images)
[Gururangan et al., 2018, Poliak et al., 2018]: good performance on SNLI based on the hypothesis alone

- This is a result of the annotation procedure
  - Negation (not, never, nobody) is correlated with *contradiction*
  - ...and “cat” as well (many dog images)
  - Generic words (animal, instrument) are correlated with *entailment*
Annotation Artifacts in SNLI

- [Gururangan et al., 2018, Poliak et al., 2018]: good performance on SNLI based on the hypothesis alone
- This is a result of the annotation procedure
  - Negation (not, never, nobody) is correlated with contradiction
  - ...and “cat” as well (many dog images)
  - Generic words (animal, instrument) are correlated with entailment
  - Sentence length: entailment < contradiction < neutral
Do neural NLI models implicitly learn lexical semantic relations?
New Test Set

- We constructed a new test set to answer this question
New Test Set

- We constructed a new test set to answer this question
- **Premise**: sentences from the SNLI training set
New Test Set

- We constructed a new test set to answer this question

- **Premise**: sentences from the SNLI training set

- **Hypothesis**: 
  - Replacing a single term $w$ in the premise with a related term $w'$
New Test Set

- We constructed a new test set to answer this question
- **Premise**: sentences from the SNLI training set
- **Hypothesis**:
  - Replacing a single term $w$ in the premise with a related term $w'$
  - $w'$ is in the SNLI vocabulary and in pre-trained embeddings
New Test Set

- We constructed a new test set to answer this question
- **Premise**: sentences from the SNLI training set
- **Hypothesis**:
  - Replacing a single term \( w \) in the premise with a related term \( w' \)
  - \( w' \) is in the SNLI vocabulary and in pre-trained embeddings
  - Crowdsourcing labels (mostly contradictions!)
New Test Set

- We constructed a new test set to answer this question
- **Premise**: sentences from the SNLI training set
- **Hypothesis**:
  - Replacing a single term $w$ in the premise with a related term $w'$
  - $w'$ is in the SNLI vocabulary and in pre-trained embeddings
  - Crowdsourcing labels (mostly contradictions!)

### Contradiction

- The man is holding a **saxophone** $\rightarrow$ The man is holding an **electric guitar**
New Test Set

- We constructed a new test set to answer this question
- **Premise**: sentences from the SNLI training set
- **Hypothesis**:
  - Replacing a single term \( w \) in the premise with a related term \( w' \)
  - \( w' \) is in the SNLI vocabulary and in pre-trained embeddings
  - Crowdsourcing labels (mostly contradictions!)

**Contradiction**

The man is holding a **saxophone** \( \rightarrow \) The man is holding an **electric guitar**

**Entailment**

A little girl is very **sad** \( \rightarrow \) A little girl is very **unhappy**
New Test Set

- We constructed a new test set to answer this question
- **Premise**: sentences from the SNLI training set
- **Hypothesis**:
  - Replacing a single term $w$ in the premise with a related term $w'$
  - $w'$ is in the SNLI vocabulary and in pre-trained embeddings
  - Crowdsourcing labels (mostly contradictions!)

### Contradiction

The man is holding a **saxophone** → The man is holding an **electric guitar**

### Entailment

A little girl is very **sad** → A little girl is very **unhappy**

### Neutral

A couple drinking **wine** → A couple drinking **champagne**
Evaluation Setting

- 3 representative models:
  - Residual-Stacked-Encoder [Nie and Bansal, 2017]
  - ESIM (Enhanced Sequential Inference Model) [Chen et al., 2017]
  - Decomposable Attention [Parikh et al., 2016]
Evaluation Setting

- 3 representative models:
  - Residual-Stacked-Encoder [Nie and Bansal, 2017]
  - ESIM (Enhanced Sequential Inference Model) [Chen et al., 2017]
  - Decomposable Attention [Parikh et al., 2016]

- Train on SNLI training set, test on the original & new test set
  - In the paper: enhancing with additional existing datasets
Results

Can neural NLI models recognize lexical inferences?

Dramatic drop in performance across models.
The test set is solvable using WordNet.
What do neural NLI models learn with respect to lexical semantic relations?
Analysis 1: Word Similarity

- Models err on contradicting word-pairs with similar embeddings
  - *A man starts his day in India → A man starts his day in Malaysia*
Analysis 1: Word Similarity

- Models err on contradicting word-pairs with similar embeddings
  - A man starts his day in India → A man starts his day in Malaysia
- Especially for fixed word embeddings

![Decomposable Attention Accuracy](chart)

- Cosine Similarity of (word, replacement)
  - 0.5-0.6: 46.2
  - 0.6-0.7: 42.3
  - 0.7-0.8: 37.5
  - 0.8-0.9: 29.7
  - 0.9-1.0: 20.2
Analysis 2: Frequency in Training

- Tuning embeddings may associate specific \((word, replacement)\) pairs to a label, e.g. \((man, woman) \rightarrow \text{contradiction}\)
Analysis 2: Frequency in Training

- Tuning embeddings may associate specific \((\text{word, replacement})\) pairs to a label, e.g. \((\text{man, woman}) \rightarrow \text{contradiction}\)
- Accuracy increases with frequency in training set

![Bar chart showing the accuracy of ESIM model with different frequencies of \((\text{word, replacement})\) pairs in contradiction training examples. Frequency ranges from 0 to 100+. The accuracy increases from 40.2% to 98.5% as the frequency increases.]
Breaking NLI
Recap

- New NLI test set that evaluates systems’ ability to make inferences that require *very simple* lexical knowledge
Breaking NLI

Recap

- New NLI test set that evaluates systems’ ability to make inferences that require *very simple* lexical knowledge

- SOTA systems perform poorly on the test set
Breaking NLI
Recap

- New NLI test set that evaluates systems’ ability to make inferences that require *very simple* lexical knowledge
- SOTA systems perform poorly on the test set
- Systems are limited in their generalization ability
**Breaking NLI**

**Recap**

- New NLI test set that evaluates systems’ ability to make inferences that require *very simple* lexical knowledge

- SOTA systems perform poorly on the test set

- Systems are limited in their generalization ability

**Related Work:**

- “Stress tests” [Naik et al., 2018]: similar findings on a broader range of linguistic phenomena
Breaking NLI

Recap

- New NLI test set that evaluates systems’ ability to make inferences that require very simple lexical knowledge
- SOTA systems perform poorly on the test set
- Systems are limited in their generalization ability

Related Work:
- “Stress tests” [Naik et al., 2018]: similar findings on a broader range of linguistic phenomena
- Inference with single word differences: [Pavlick and Callison-Burch, 2016, Kalouli et al., 2018]
But current LM-based NLI models address entailment-related phenomena better, no?
Pre-trained LM based NLI models

[CLS] Premise [SEP] Hypothesis
Diversify Your Datasets: Analyzing Generalization via Controlled Variance in Adversarial Datasets

Ohad Rozen, Vered Shwartz, Roee Aharoni, and Ido Dagan

Bar-Ilan University
CoNLL 2019
MultiNLI [Williams et al., 2018]

- Collected like SNLI (existing premises, generated hypotheses)
MultiNLI [Williams et al., 2018]

- Collected like SNLI (existing premises, generated hypotheses)
- Multiple genres
MultiNLI [Williams et al., 2018]

- Collected like SNLI (existing premises, generated hypotheses)
- Multiple genres
- Mismatched evaluation (not in our focus)
Probing → Inoculation

- **Probing**: does the representation capture a certain property? [Glockner et al., 2018, Naik et al., 2018]
Probing → Inoculation

- **Probing**: does the representation *capture* a certain property? [Glockner et al., 2018, Naik et al., 2018]

- **Inoculation** [Liu et al., 2019]: can the representation *learn* a certain property?
Has the model learned a general notion of the property, or does it overfit to the specific dataset?
Testing Generalization Capacity

Methodology

1. Split the challenge dataset to different variations across the dimension in focus.

- (A) Numerical Reasoning
  - (A-1) Range: 100 - 200
  - (A-2) Range: 400 - 500
Testing Generalization Capacity

Methodology

2. Fine-tune on one set and test on another.

Fine-Tune

(A-1)
Range: 100 - 200

Test

(A-2)
Range: 400 - 500
Testing Generalization Capacity

Methodology

Fine-Tune

(A-1)
Range: 100 - 200

Test

(A-2)
Range: 400 - 500

Failure
Poor generalization capacity
Testing Generalization Capacity
Methodology

Fine-Tune
(A-1)
Range: 100 - 200

Test
(A-2)
Range: 400 - 500

Failure
Poor generalization capacity

Success
Good generalization capacity
Challenge Datasets Generation

Templates from MultiNLI sentences →
Ask crowdsourcing workers to rephrase spans.

1. **Extracted Premise**: [The Citigroup deal], [from beginning to end], [took] less than 5 [weeks].
2. **Premise Template**: ARG1, ARG2, ARG3 RELATION NUM ARG4.
   **Hypothesis Template (Ent.)**: ARG1, ARG2, ARG3 more than NUM-smaller ARG4.
3. **Gen. Premise**: [My marriage], [despite much frustration], [lasted] more than 7 [years].
   **Gen. Hypothesis (Ent.)**: [My marriage], [despite much frustration], [lasted] more than 2 [years].

1000s of different training examples with similar syntax from 1 original sentence
Phenomena
Dative Alternation

Premise:   *I baked my mom a cake*
Hypothesis 1: *I baked a cake for my mom*
Label: Entailment
Phenomena
Numeric Reasoning

Premise: I see 260 coins in the bucket.
Hypothesis: I see more than 232 coins in the bucket.
Label: Entailment
Challenge Datasets Generation
Diversity Axes

1. Syntax complexity - simple / medium / complex
   - Sentence length
   - Phenomenon depth in parse tree
Challenge Datasets Generation

Diversity Axes

1. Syntax complexity - simple / medium / complex
   - Sentence length
   - Phenomenon depth in parse tree

2. Lexical variability
   - Dative verb
   - Number range
Inoculation tells part of the story...
Dative Alternation
Not sensitive to lexical variability
...our data tells the other part

Dative Alternation
Not sensitive to *lexical* variability
But generalizes only from complex to simple *syntax*
...our data tells the other part

**Dative Alternation**
Not sensitive to *lexical* variability
But generalizes only from complex to simple *syntax*

**Numeric Reasoning**
Not sensitive to *syntax*
...our data tells the other part

**Dative Alternation**
Not sensitive to *lexical* variability
But generalizes only from complex to simple *syntax*

**Numeric Reasoning**
Not sensitive to *syntax*
But doesn’t generalize across *number ranges*
Diversify your Datasets

Recap

- Simple methodology to test model generalization of a specific learned phenomenon
Diversify your Datasets

Recap

- Simple methodology to test model generalization of a specific learned phenomenon
- NLI-BERT fails to generalize dative alternation and numeric reasoning
Diversify your Datasets

Recap

- Simple methodology to test model generalization of a specific learned phenomenon
- NLI-BERT fails to generalize dative alternation and numeric reasoning
- Fine-tuning on the phenomenon-specific data may decrease the main task performance (also in [Richardson et al., 2020]).
Real-world examples: partial entailments

$S_1$: Amazon To Acquire Whole Foods Market For $13.7$ Billion

$S_2$: Amazon is buying Whole Foods for almost $14$ billion in cash
Real-world examples: partial entailments

$S_1$: Researchers have discovered wreckage of the lost warship, the USS Indianapolis after 72 years

$S_2$: Wreckage of missing WWII ship found in Pacific Ocean
NLI

Coreference

Breaking

[Shwartz et al., 2017]
[Barhom et al., 2019]
Revisiting Joint Modeling of Cross-document Entity and Event Coreference Resolution

Shany Barhom\textsuperscript{1}, Vered Shwartz\textsuperscript{1}, Alon Eirew\textsuperscript{2}, Michael Bugert\textsuperscript{3}, Nils Reimers\textsuperscript{3}, and Ido Dagan\textsuperscript{1}

\textsuperscript{1}Bar-Ilan University \quad \textsuperscript{2}Intel AI Lab \quad \textsuperscript{3}TU Darmstadt

ACL 2019
Cross-document Coreference Resolution

---

Tara Reid has entered a rehab center…

...She checked into the facility today…
Cross-document Coreference Resolution

Doc #1

Tara Reid has entered a rehab center…

...She checked into the facility today...
Tara Reid has entered a rehab center…

...She checked into the facility today…
Cross-document Coreference Resolution

Doc #1
Tara Reid has entered a rehab center...
...She checked into the facility today...

Doc #2
...the American Pie star headed to a Malibu treatment facility on Tuesday...
Our Research Focus

- Entity and Event Coreference are closely interdependent
Our Research Focus

- Entity and Event Coreference are closely interdependent

1. Tara Reid has entered a rehab center.
2. The American Pie star headed to a Malibu treatment facility on Tuesday.
Our Research Focus

- Entity and Event Coreference are closely interdependent

1. Tara Reid has entered a rehab center.

2. The American Pie star headed to a Malibu treatment facility on Tuesday.
Our Research Focus

- Entity and Event Coreference are closely interdependent

1. Tara Reid has entered a rehab center.
2. The American Pie star headed to a Malibu treatment facility on Tuesday.
Our Research Focus

- Entity and Event Coreference are closely interdependent

1. Tara Reid has entered a rehab center.

2. The American Pie star headed to a Malibu treatment facility on Tuesday.
Our Research Focus

- Entity and Event Coreference are closely interdependent

1. Tara Reid has entered a rehab center.
2. The American Pie star headed to a Malibu treatment facility on Tuesday.
Our Research Focus

- Entity and Event Coreference are closely interdependent

1. Tara Reid has entered a rehab center.

2. The American Pie star headed to a Malibu treatment facility on Tuesday.
Our Research Focus

- Entity and Event Coreference are closely interdependent

1. Tara Reid has entered a rehab center.
2. The American Pie star headed to a Malibu treatment facility on Tuesday.
Our Research Focus

- Entity and Event Coreference are closely interdependent

1. Tara Reid has entered a rehab center.

2. The American Pie star headed to a Malibu treatment facility on Tuesday.
Our Research Focus

- Entity and Event Coreference are closely interdependent - calls for a joint approach
Our Research Focus

- Entity and Event Coreference are closely interdependent - calls for a joint approach

- Only single such prior work (Lee et al., 2012)
Our Research Focus

- Entity and Event Coreference are closely interdependent - calls for a joint approach

- Only single such prior work (Lee et al., 2012)

- We revisit the joint resolution approach, suggesting new neural models to address it
  - Achieving new SOTA
Related Work
Within-document Coreference
Within-document Entity Coreference

Lee et al., 2018
Lee et al., 2017
Clark and Manning, 2016
Wiseman et al. 2016
Wiseman et al., 2015
Martschat and Strube, 2015
Durrett and Klein, 2013

CoNLL- 2012 PreCo
Within-document Event Coreference

Choubey and Huang 2017b
Choubey and Huang, 2017a
Lu et al. 2016
Krause et al., 2016
Lu and Ng, 2016
Liu et al., 2014
Chen et al., 2009
....

TAC-KBP
ACE
Cross-document Event Coreference

Kenyon-Dean et al., 2018
Choubey and Huang, 2017
Cybulska and Vossen, 2015
Yang et al., 2015
Lee et al., 2012

ECB+  GVC
Cross-document Entity Coreference

Dutta and Weikum, 2015
Singh et al., 2015
Lee et al., 2012
Rao et al., 2010

WePS-2 NYT
Common Approach - Lexical Similarity between Arguments

1. **Tara Reid** has entered **a rehab center**

2. **The American Pie star** headed **to a Malibu treatment facility** on Tuesday
Common Approach - Lexical Similarity between Arguments

1. Tara Reid has entered a rehab center

2. The American Pie star headed to a Malibu treatment facility on Tuesday
Common Approach - Lexical Similarity between Arguments

1. Tara Reid has entered a rehab center

2. The American Pie star headed to a Malibu treatment facility on Tuesday

Is that good enough? 😐
Joint Entity and Event Coreference

- Lee et al., (2012) introduced a system that models entity and event coreference jointly
Joint Entity and Event Coreference

- Lee et al., (2012) introduced a system that models entity and event coreference jointly

- Iterative method that constructs clusters of entity and event mentions
Joint Entity and Event Coreference

- Lee et al., (2012) introduced a system that models entity and event coreference jointly
- Iterative method that constructs clusters of entity and event mentions
- Linear regression to model cluster merge operations, based on discrete features
Joint Entity and Event Coreference

- Lee et al., (2012) introduced a system that models entity and event coreference jointly

- Iterative method that constructs clusters of entity and event mentions

- Linear regression to model cluster merge operations, based on discrete features

- We revisit the joint approach, suggesting a neural models to address it
Algorithm Flow - Alternating Between Entity and Event Clustering
Algorithm Flow - Alternating Between Entity and Event Clustering
Algorithm Flow - Alternating Between Entity and Event Clustering

Entity Clustering

Event Clustering

Multiple event cluster merges
Algorithm Flow - Alternating Between Entity and Event Clustering
Cluster Merging Score

- Hierarchical clustering requires a cluster pair merging score
- **Average link:** average all mention pair scores across the two candidate clusters

\[ S_{cp}(c_i, c_j) = \frac{1}{|c_i| \cdot |c_j|} \cdot \sum_{m_i \in c_i} \sum_{m_j \in c_j} S(m_i, m_j) \]
Cluster Merging Score

- Hierarchical clustering requires a cluster pair merging score

- **Average link:** average all mention pair scores across the two candidate clusters

\[
S_{cp}(c_i, c_j) = \frac{1}{|c_i| \cdot |c_j|} \cdot \sum_{m_i \in c_i} \sum_{m_j \in c_j} S(m_i, m_j)
\]
1. Tara Reid has entered a rehab center.
2. The American Pie star headed to a Malibu treatment facility on Tuesday.
1. Tara Reid has entered a rehab center.
2. The American Pie star headed to a Malibu treatment facility on Tuesday.
1. Tara Reid has entered a rehab center
2. The American Pie star headed to a Malibu treatment facility on Tuesday
Mention-Pair Representation

1. Tara Reid has entered a rehab center
2. The American Pie star headed to a Malibu treatment facility on Tuesday
Mention-Pair Representation

1. Tara Reid has entered a rehab center
2. The American Pie star headed to a Malibu treatment facility on Tuesday
Mention-Pair Representation

1. Tara Reid has entered a rehab center
2. The American Pie star headed to a Malibu treatment facility on Tuesday

Predicate-Argument Vectors

Average Pooling

Mention Representations

Context Span Arg0 Arg1 Location Time

Tara Reid
The Actress
Reid
The American Pie star
Mention-Pair Representation

1. Tara Reid has entered a rehab center
2. The American Pie star headed to a Malibu treatment facility on Tuesday
1. Tara Reid has entered a rehab center
2. The American Pie star headed to a Malibu treatment facility on Tuesday
1. Tara Reid has entered a rehab center
2. The American Pie star headed to a Malibu treatment facility on Tuesday
1. Tara Reid has entered a rehab center
2. The American Pie star headed to a Malibu treatment facility on Tuesday
Mention-Pair Representation

1. Tara Reid has entered a rehab center
2. The American Pie star headed to a Malibu treatment facility on Tuesday

High similarity between argument vectors!
1. Tara Reid has entered a rehab center
2. The American Pie star headed to a Malibu treatment facility on Tuesday
Mention-Pair Scorer

$$S_{cp}(c_i, c_j) = \frac{1}{|c_i| \cdot |c_j|} \cdot \sum_{m_i \in c_i} \sum_{m_j \in c_j} S(m_i, m_j)$$
Training

- We train two distinct pairwise scorers (one for entities and one for events)
Training

- We train two distinct pairwise scorers (one for entities and one for events)
- The training procedure simulates the inference step
  - Allows the models to be trained on various predicted clustering configurations
Training

- We train two distinct pairwise scorers (one for entities and one for events)

- The training procedure simulates the inference step
  - Allows the models to be trained on various predicted clustering configurations

- **Training examples:** all mention pairs that belong to different clusters in the current clustering configuration
Training

- We train two distinct pairwise scorers (one for entities and one for events)

- The training procedure simulates the inference step
  - Allows the models to be trained on various predicted clustering configurations

- **Training examples:** all mention pairs that belong to different clusters in the current clustering configuration

- Scorers are repeatedly trained and then used for clusters merging
Experiments
Dataset

- ECB+ (Event-Coreference-Bank; Cybulska and Vossen, 2014).
  - Within- and cross-document coreference annotations for entities and events
  - ~1000 documents, clustered into 43 topics, discussing different seminal events
Dataset

- ECB+ (Event-Coreference-Bank; Cybulska and Vossen, 2014).
  - Within- and cross-document coreference annotations for both entities and events
  - ~1000 documents, clustered into 43 topics, discussing different seminal events

Topic 1: A celebrity enters into rehab

...Tara Reid finally checks into rehab...

...Actress Tara Reid entered well-known Malibu rehab center ...

...Lindsay Lohan checks into rehab...

...The NYDN says that LiLo eventually made it to Morningside Recovery...
Evaluation Setup

- We follow Cybulskak and Vossen (2015) and Kenyon-Dean et al., (2018)
  - Corpus’s subset which has been validated for correctness
  - Use the gold mentions during evaluation

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Validation</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># Topics</td>
<td>25</td>
<td>8</td>
<td>10</td>
<td>43</td>
</tr>
<tr>
<td># Sub-topics</td>
<td>50</td>
<td>16</td>
<td>20</td>
<td>86</td>
</tr>
<tr>
<td># Documents</td>
<td>574</td>
<td>196</td>
<td>206</td>
<td>976</td>
</tr>
<tr>
<td># Sentences</td>
<td>1037</td>
<td>346</td>
<td>457</td>
<td>1840</td>
</tr>
<tr>
<td># Event mentions</td>
<td>3808</td>
<td>1245</td>
<td>1780</td>
<td>6833</td>
</tr>
<tr>
<td># Entity mentions</td>
<td>4758</td>
<td>1476</td>
<td>2055</td>
<td>8289</td>
</tr>
<tr>
<td># Event chains</td>
<td>1527</td>
<td>409</td>
<td>805</td>
<td>2741</td>
</tr>
<tr>
<td># Entity chains</td>
<td>1286</td>
<td>330</td>
<td>608</td>
<td>2224</td>
</tr>
</tbody>
</table>
Evaluation Setup

- We follow Cybulska and Vossen (2015) and Kenyon-Dean et al., (2018)
  - Corpus’s subset which has been validated for correctness
  - Use the gold mentions during evaluation

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Validation</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># Topics</td>
<td>25</td>
<td>8</td>
<td>10</td>
<td>43</td>
</tr>
<tr>
<td># Sub-topics</td>
<td>50</td>
<td>16</td>
<td>20</td>
<td>86</td>
</tr>
<tr>
<td># Documents</td>
<td>574</td>
<td>196</td>
<td>206</td>
<td>976</td>
</tr>
<tr>
<td># Sentences</td>
<td>1037</td>
<td>346</td>
<td>457</td>
<td>1840</td>
</tr>
<tr>
<td># Event mentions</td>
<td>3808</td>
<td>1245</td>
<td>1780</td>
<td>6833</td>
</tr>
<tr>
<td># Entity mentions</td>
<td>4758</td>
<td>1476</td>
<td>2055</td>
<td>8289</td>
</tr>
<tr>
<td># Event chains</td>
<td>1527</td>
<td>409</td>
<td>805</td>
<td>2741</td>
</tr>
<tr>
<td># Entity chains</td>
<td>1286</td>
<td>330</td>
<td>608</td>
<td>2224</td>
</tr>
</tbody>
</table>
Coreference Results

Event Coreference Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Test CoNLL F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kenyon-Dean et al., (2018)</td>
<td>69</td>
</tr>
<tr>
<td>Cybulska and Vossen (2015)</td>
<td>73</td>
</tr>
<tr>
<td>CLUSTER + Kenyon-Dean et al., (2018)</td>
<td>73.6</td>
</tr>
<tr>
<td>CLUSTER+ LEMMA</td>
<td>76.5</td>
</tr>
<tr>
<td>DISJOINT</td>
<td>78.5</td>
</tr>
<tr>
<td>JOINT</td>
<td>79.5</td>
</tr>
</tbody>
</table>
Coreference Results

Entity Coreference Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Test CoNLL F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLUSTER+ LEMMA</td>
<td>67.4</td>
</tr>
<tr>
<td>DISJOINT</td>
<td>70</td>
</tr>
<tr>
<td>JOINT</td>
<td>71.2</td>
</tr>
</tbody>
</table>
Coreference Results

First entity coreference results on ECB+
Correct Model’s Decisions

Entity coreference:
- [WWDC, San Francisco gathering's, conference]
- [West Papua, region, in remote eastern Indonesia]
- [Matt Smith, actor]

Event coreference:
- [launches, unveiled]
- [rattled, struck, hit]
- [acquires, buys, purchase]
Wrong Model’s Decisions

Entity coreference:
- [next generation of MacBook Pro, MacBook Pro]
- [five people, four people]
- [Wednesday, on Monday]

Event coreference:
- [recorded, occurred]
- [sales, acquisition]
- [gone official, go ahead]
Wrong Model’s Decisions

Entity coreference:
- [next generation of MacBook Pro, MacBook Pro]
- [five people, four people]
- [Wednesday, on Monday]

Event coreference:
- [recorded, occurred]
- [sales, acquisition]
- [gone official, go ahead]

Paraphrasing vs. Relatedness
Wrong Model's Decisions

Entity coreference:
- [next generation of MacBook Pro, MacBook Pro]
- [five people, four people]
- [Wednesday, on Monday]

Event coreference:
- [recorded, occurred]
- [sales, acquisition]
- [gone official, go ahead]
Error Analysis

---

**Event Errors**

- Same head lemma: 46%
- Predicate argument extraction error: 12%
- Entity coreference error: 10%
- Similar context: 4%
- Annotation error: 4%
- Other: 4%
- Partial argument coreference: 2%

**Entity Errors**

- Same head lemma: 44%
- Predicate argument extraction error: 4%
- Event coreference error: 6%
- Similar context: 12%
- Annotation error: 8%
- Other: 8%
- Within-document coreference error: 8%
Joint Event and Entity Coreference Resolution

Recap

- Cross-document coreference is drastically under-explored
Joint Event and Entity Coreference Resolution

Recap

- Cross-document coreference is drastically under-explored
- A simple joint approach with state-of-the-art results on ECB+
Joint Event and Entity Coreference Resolution

Recap

- Cross-document coreference is drastically under-explored
- A simple joint approach with state-of-the-art results on ECB+
- Still a long way to go!
Acquiring Predicate Paraphrases from News Tweets

Vered Shwartz, Gabriel Stanovsky, and Ido Dagan

Bar-Ilan University

*SEM 2017
Acquiring Predicate Paraphrases from News Tweets

Binary verbal predicate paraphrases
Extracted from Twitter
Ever-growing resource: currently around 5.2M paraphrases

Available at https://github.com/vered1986/Chirps
Assumptions

- **Main assumption:** Redundant news headlines of the same event are likely to describe it with different words [Shinyama et al., 2002, Barzilay and Lee, 2003].
Main assumption: redundant news headlines of the same event are likely to describe it with different words [Shinyama et al., 2002, Barzilay and Lee, 2003].

This work: propositions extracted from tweets discussing news events, published on the same day, that agree on their arguments, are predicate paraphrases.
Resource Collection

- Collect News Tweets
- Extract Propositions
- Generate Paraphrase Instances
- Generate Types
- Resource Release

Query the Twitter Search API for news tweets in English

Amazon is buying Whole Foods in $13.7B
Amazon to acquire Whole Foods Market in deal valued at nearly $14 billion

...
Extract propositions from tweets using PropS [Stanovsky et al., 2016]

Get binary verbal predicate templates, and apply argument reduction [Stanovsky and Dagan, 2016]

[Amazon] buy [Whole Foods]
[Amazon] acquire [Whole Foods Market]

...
We consider two predicates as paraphrases if:

1. They appear on the same day.
2. Each of their arguments aligns with a unique argument in the other predicate.

Two levels of argument matching: **strict** (exact match / short edit distance) and **loose** (partial token matching / WordNet synonyms)
Resource Collection

Heuristic score for a predicate paraphrase type:

\[ p_1 = [a]_0 \text{ buy } [a]_1, \quad p_2 = [a]_0 \text{ acquire } [a]_1 \]

\[
s(p_1, p_2) = \text{count}(p_1, p_2) \cdot \left(1 + \frac{\text{days}(p_1, p_2)}{N}\right)
\]

- \(\text{count}(p_1, p_2)\) assigns high scores for frequent paraphrases
- \(N\) - number of days since the resource collection begun
- \(\frac{\text{days}(p_1, p_2)}{N}\) eliminates noise from two arguments participating in different events on the same day

1) Last year when Chuck Berry turned 90; 2) Chuck Berry dies at 90
We release our resource daily, with two files:

- **Instances**: predicates, arguments and tweet IDs.
- **Types**: predicate paraphrase pair types ranked in a descending order according to the heuristic accuracy score.
Chirps Recap

- Using event coreference to extract paraphrases
Chirps
Recap

- Using event coreference to extract paraphrases
- Complementary to other paraphrasing resources
Using event coreference to extract paraphrases

Complementary to other paraphrasing resources

Useful resource for paraphrasing, event coreference, NLI
Thank you!

Questions?

@VeredShwartz  vereds@allenai.org
References I


References II


References III


