Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation

EMNLP 2018

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, Benjamin Van Durme
Natural Language Inference

Premise: *The brown cat ran*

Hypothesis: *The animal moved*
Natural Language Inference

Premise: *The brown cat ran*

Hypothesis: *The animal moved*

entailed  not-entailed
Natural Language Inference

Premise: *The brown cat ran*

Hypothesis: *The animal moved*

entailed

not-entailed

Multiple labeling schemas
Natural Language Inference

Premise: *The brown cat ran*

Hypothesis: *The animal moved*

- entailed
- not-entailed
- entailment
- neutral
- contradiction

Multiple labeling schemas
Natural Language Inference

Premise: *The brown cat ran*

Hypothesis: *The animal moved*

entailed  not-entailed
Natural Language Inference

Premise: The brown cat ran

Hypothesis: The animal moved

entailed not-entailed
Natural Language Inference

Premise: *The brown cat ran*

Hypothesis: *The animal moved*

entailed

not-entailed
Natural Language Inference

Premise: *The brown cat ran*

Hypothesis: *The animal moved*

entailed  not-entailed
Why NLI as an NLP task?
Evaluation & Probing models
Historically
Historically

\textbf{FraCas:}

(Cooper et al., 1996)
Historically

_FraCas_ determine whether a model performs distinct types of reasoning

(Cooper et al., 1996)
Historically

**FraCas**: determine whether a model performs distinct types of reasoning  
(Cooper et al., 1996)

**Pascal RTE**:  
(Dagan et al., 2006)
Historically

**FraCas:** determine whether a model performs distinct types of reasoning

(Cooper et al., 1996)

**Pascal RTE:** “a generic evaluation framework” to compare models for distinct downstream tasks

(Dagan et al., 2006)
More recent
More recent

**SNLI & Multi-NLI:**

(Bowman et. al. 2015; Williams et. al. 2018)
More recent

**SNLI & Multi-NLI**: large scale datasets

(Bowman et. al. 2015; Williams et. al. 2018)
More recent

**SNLI & Multi-NLI**: large scale datasets

(Bowman et. al. 2015; Williams et. al. 2018)

Evaluate sentence representations

(Rep Eval 2017 Shared Task - Nangia et. al. 2017)
More recent

**SNLI & Multi-NLI:** large scale datasets

(Bowman et. al. 2015; Williams et. al. 2018)

Evaluate sentence representations

(Rep Eval 2017 Shared Task - Nangia et. al. 2017)

Training to improve models for downstream tasks

(Guo et. al. 2018)
Prior Dataset Characteristics
Prior Dataset Characteristics

NLU Insights
Prior Dataset Characteristics

NLU Insights

Generation Methods
Prior Dataset Characteristics

NLU Insights

Generation Methods

Small Probing Sets
Characteristic 1: NLU Insights

Understanding our models’ reasoning capabilities?
Characteristic 1: NLU Insights

<table>
<thead>
<tr>
<th>Model</th>
<th>Type Description</th>
<th>F1</th>
<th>EM</th>
<th>Exact Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jianpeng Cheng et al. '16</td>
<td>450D LSTM w/ deep attention fusion</td>
<td>3.4m</td>
<td>88.5</td>
<td>86.3</td>
</tr>
<tr>
<td>Parikh et al. '16</td>
<td>200D decomposable attention model</td>
<td>380k</td>
<td>89.5</td>
<td>86.3</td>
</tr>
<tr>
<td>Parikh et al. '16</td>
<td>200D decomposable attention model with intra-sentence attention</td>
<td>580k</td>
<td>90.5</td>
<td>86.8</td>
</tr>
<tr>
<td>Munkhdalai &amp; Yu '16b</td>
<td>300D Full tree matching NTI-SLSTM-LSTM w/ global attention</td>
<td>3.2m</td>
<td>88.5</td>
<td>87.3</td>
</tr>
<tr>
<td>Zhiguo Wang et al. '17</td>
<td>BiMMP</td>
<td>1.6m</td>
<td>90.9</td>
<td>87.5</td>
</tr>
<tr>
<td>Leli Sha et al. '16</td>
<td>300D re-read LSTM</td>
<td>2.0m</td>
<td>90.7</td>
<td>87.5</td>
</tr>
<tr>
<td>Yichen Gong et al. '17</td>
<td>448D Densely Interactive Inference Network (DIIN, code)</td>
<td>4.4m</td>
<td>91.2</td>
<td>88.0</td>
</tr>
<tr>
<td>McCann et al. '17</td>
<td>Biattentive Classification Network + CoVe + Char</td>
<td>22m</td>
<td>88.5</td>
<td>88.1</td>
</tr>
<tr>
<td>Chuanqi Tan et al. '18</td>
<td>150D Multiway Attention Network</td>
<td>14m</td>
<td>94.5</td>
<td>88.3</td>
</tr>
<tr>
<td>Xiaodong Liu et al. '18</td>
<td>Stochastic Answer Network</td>
<td>3.5m</td>
<td>93.3</td>
<td>88.5</td>
</tr>
<tr>
<td>Ghaeini et al. '18</td>
<td>450D DR-BiLSTM</td>
<td>7.5m</td>
<td>94.1</td>
<td>88.5</td>
</tr>
<tr>
<td>Yi Tay et al. '18</td>
<td>300D CAFE</td>
<td>4.7m</td>
<td>89.8</td>
<td>88.5</td>
</tr>
<tr>
<td>Qian Chen et al. '17</td>
<td>KIM</td>
<td>4.3m</td>
<td>94.1</td>
<td>88.6</td>
</tr>
<tr>
<td>Qian Chen et al. '16</td>
<td>600D EISIM + 300D Syntactic TreeLSTM (code)</td>
<td>7.7m</td>
<td>93.5</td>
<td>88.6</td>
</tr>
<tr>
<td>Peters et al. '18</td>
<td>EISIM + ELMo</td>
<td>8.0m</td>
<td>91.6</td>
<td>88.7</td>
</tr>
<tr>
<td>Boyuan Pan et al. '18</td>
<td>300D DMAN</td>
<td>9.2m</td>
<td>95.4</td>
<td>88.8</td>
</tr>
<tr>
<td>Zhiguo Wang et al. '17</td>
<td>BiMMP Ensemble</td>
<td>6.4m</td>
<td>93.2</td>
<td>88.8</td>
</tr>
<tr>
<td>Yichen Gong et al. '17</td>
<td>448D Densely Interactive Inference Network (DIIN, code) Ensemble</td>
<td>17m</td>
<td>92.3</td>
<td>88.9</td>
</tr>
<tr>
<td>Seonhoon Kim et al. '18</td>
<td>Densely-Connected Recurrent and Co-Attentive Network</td>
<td>6.7m</td>
<td>93.1</td>
<td>88.9</td>
</tr>
<tr>
<td>Zhuosheng Zhang et al. '18</td>
<td>SLRC</td>
<td>6.1m</td>
<td>89.1</td>
<td>89.1</td>
</tr>
<tr>
<td>Qian Chen et al. '17</td>
<td>KIM Ensemble</td>
<td>43m</td>
<td>93.6</td>
<td>89.1</td>
</tr>
<tr>
<td>Ghaeini et al. '18</td>
<td>450D DR-BiLSTM Ensemble</td>
<td>45m</td>
<td>94.8</td>
<td>89.3</td>
</tr>
<tr>
<td>Peters et al. '18</td>
<td>EISIM + ELMo Ensemble</td>
<td>40m</td>
<td>92.1</td>
<td>89.3</td>
</tr>
<tr>
<td>Yi Tay et al. '18</td>
<td>300D CAFE Ensemble</td>
<td>17.5m</td>
<td>92.5</td>
<td>89.3</td>
</tr>
<tr>
<td>Chuanqi Tan et al. '18</td>
<td>150D Multiway Attention Network Ensemble</td>
<td>58m</td>
<td>95.5</td>
<td>89.4</td>
</tr>
<tr>
<td>Boyuan Pan et al. '18</td>
<td>300D DMAN Ensemble</td>
<td>79m</td>
<td>96.1</td>
<td>89.6</td>
</tr>
<tr>
<td>Radford et al. '18</td>
<td>Fine-Tuned LM-Pretrained Transformer</td>
<td>85m</td>
<td>96.6</td>
<td>89.9</td>
</tr>
<tr>
<td>Seonhoon Kim et al. '18</td>
<td>Densely-Connected Recurrent and Co-Attentive Network Ensemble</td>
<td>53.3m</td>
<td>95.0</td>
<td>90.1</td>
</tr>
</tbody>
</table>
Characteristic 2: Generation Methods
Characteristic 2: Generation Methods

Expensive
Characteristic 2: Generation Methods

Expensive

Leads to biases:
Characteristic 2: Generation Methods

Expensive

Leads to biases:

Stereotypical

(Rudinger et. al. 2017)
Characteristic 2: Generation Methods

Expensive

Leads to biases:

Stereotypical
(Rudinger et. al. 2017)

Class-based Statistical Irregularities
(Tsuchiya, 2018; Gururangan et al., 2018; Poliak et al., 2018)
Characteristic 3: Small Probing Sets
Characteristic 3: Small Probing Sets

FraCas is too small

Training neural network on 300 examples
Outline

● Introduction

● The DNC: Diverse NLI Collection

● Constructing the DNC

● Experiments & Results
The DNC
The DNC

Diverse Natural Language Inference Collection
The DNC

Diverse Natural Language Inference Collection

Large scale collection of diverse NLI problems
The DNC

Diverse Natural Language Inference Collection

Large scale collection of diverse NLI problems

Convert 7 semantic phenomena into NLI from 13 existing datasets
## The DNC - Examples

<table>
<thead>
<tr>
<th>Event</th>
<th>Factuality</th>
<th>Puns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find him before he finds the dog food</td>
<td>The finding did not happen</td>
<td><img src="%E2%9C%93" alt="Kim heard masks have no face value" /></td>
</tr>
<tr>
<td>I’ll need to ponder</td>
<td>The pondering happened</td>
<td><img src="%E2%9C%93" alt="Kim heard a pun" /></td>
</tr>
<tr>
<td>Ward joined Tom in their native Perth</td>
<td>Ward was born in Perth</td>
<td><img src="%E2%9C%97" alt="Tod heard that thrift is better than annuity" /></td>
</tr>
<tr>
<td>Stefan had visited his son in Bulgaria</td>
<td>Stefan was born in Bulgaria</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><img src="%E2%9C%97" alt="Tod heard a pun" /></td>
</tr>
</tbody>
</table>
# The DNC

<table>
<thead>
<tr>
<th>Sem. Phenomena/Annotations</th>
<th>Dataset</th>
<th># pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event Factuality</td>
<td>Decomp (Rudinger et al., 2018b)</td>
<td>42K (41,888)</td>
</tr>
<tr>
<td></td>
<td>UW (Lee et al., 2015)</td>
<td>5K (5,094)</td>
</tr>
<tr>
<td></td>
<td>MeanTime (Minard et al., 2016)</td>
<td>.7K (738)</td>
</tr>
<tr>
<td>Named Entity Recognition</td>
<td>Groningen (Bos et al., 2017)</td>
<td>260K (261,406)</td>
</tr>
<tr>
<td></td>
<td>CoNLL (Tjong Kim Sang and De Meulder, 2003)</td>
<td>60K (59,970)</td>
</tr>
<tr>
<td>Gendered Anaphora</td>
<td>Winogender (Rudinger et al., 2018a)</td>
<td>.4K (464)</td>
</tr>
<tr>
<td>Lexicosyntactic Inference</td>
<td>VerbCorner (Hartshorne et al., 2013)</td>
<td>135K (138,648)</td>
</tr>
<tr>
<td></td>
<td>MegaVeridicality (White and Rawlins, 2018)</td>
<td>11K (11,814)</td>
</tr>
<tr>
<td></td>
<td>VerbNet (Schuler, 2005)</td>
<td>2K (1,950)</td>
</tr>
<tr>
<td>Puns</td>
<td>(Yang et al., 2015)</td>
<td>9K (9,492)</td>
</tr>
<tr>
<td></td>
<td>SemEval 2017 Task 7 (Miller et al., 2017)</td>
<td>8K (8,054)</td>
</tr>
<tr>
<td>Relationship Extraction</td>
<td>FACC1 (Gabrilovich et al., 2013)</td>
<td>30K (30,920)</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>(Kotzias et al., 2015)</td>
<td>6K (6,000)</td>
</tr>
<tr>
<td>Combined</td>
<td>Diverse NLI Collection (DNC)</td>
<td>575K (576,438)</td>
</tr>
<tr>
<td></td>
<td>SNLI (Bowman et al., 2015)</td>
<td>570K</td>
</tr>
<tr>
<td></td>
<td>Multi-NLI (Williams et al., 2017)</td>
<td>433K</td>
</tr>
</tbody>
</table>
Diverse Natural Language Inference Collection - NLI dataset that can used to evaluate how well models perform distinct types of reasoning (EMNLP 2018)  
http://decomp.io/projects/diverse-nat...

natural-language-processing  natural-language-inference  computational-semantics  emnlp2018

6 commits  1 branch  1 release  1 contributor

Branch: master  New pull request

azpoliak update README.md - inference is everything data  Latest commit 6a8beee on Sep 14

- dev
  Released DNC and updated README
  2 months ago

- test
  Released DNC and updated README
  2 months ago

- train
  Released DNC and updated README
  2 months ago

- README.md
  update README.md - inference is everything data
  a month ago

- additional_references.md
  added bibs for original datasets
  2 months ago

- inference_is_everything.zip
  included White et al’s IJCNLP 2017 recast data
  a month ago

---

DNC: Diverse Natural Language Inference Collection

Dataset associated and released as part of Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation (EMNLP 2018).
Diverse Natural Language Inference Collection - NLI dataset that can be used to evaluate how well models perform distinct types of reasoning (EMNLP 2018) [http://decomp.io/projects/diverse-nat...](http://decomp.io/projects/diverse-nat...)

natural-language-processing  natural-language-inference  computational-semantics  emnlp2018

6 commits  1 branch  1 release  1 contributor

Branch: master  New pull request

**azpoliak** update README.md - inference is everything data ...

- dev
- test
- train
- README
- additional_release.txt
- inference_is_everything.zip

README.md

**DNC: Diverse Natural Language Inference Collection**

Dataset associated and released as part of *Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation* (EMNLP 2018).
The DNC

<table>
<thead>
<tr>
<th>Sem. Phenomena/Annotations</th>
<th>Dataset</th>
<th># pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event Factuality</td>
<td>Decomp (Rudinger et al., 2018b)</td>
<td>42K (41,888)</td>
</tr>
<tr>
<td></td>
<td>UW (Lee et al., 2015)</td>
<td>5K (5,094)</td>
</tr>
<tr>
<td></td>
<td>MeanTime (Minard et al., 2016)</td>
<td>.7K (738)</td>
</tr>
<tr>
<td>Named Entity Recognition</td>
<td>Groningen (Bos et al., 2017)</td>
<td>260K (261,406)</td>
</tr>
<tr>
<td></td>
<td>CoNLL (Tjong Kim Sang and De Meulder, 2003)</td>
<td>60K (59,970)</td>
</tr>
<tr>
<td>Gendered Anaphora</td>
<td>Winogender (Rudinger et al., 2018a)</td>
<td>.4K (464)</td>
</tr>
<tr>
<td>Lexicosyntactic Inference</td>
<td>VerbCorner (Hartshorne et al., 2013)</td>
<td>135K (138,648)</td>
</tr>
<tr>
<td></td>
<td>MegaVeridicality (White and Rawlins, 2018)</td>
<td>11K (11,814)</td>
</tr>
<tr>
<td></td>
<td>VerbNet (Schuler, 2005)</td>
<td>2K (1,950)</td>
</tr>
<tr>
<td>Puns</td>
<td>(Yang et al., 2015)</td>
<td>9K (9,492)</td>
</tr>
<tr>
<td></td>
<td>SemEval 2017 Task 7 (Miller et al., 2017)</td>
<td>8K (8,054)</td>
</tr>
<tr>
<td>Relationship Extraction</td>
<td>FACC1 (Gabrilovich et al., 2013)</td>
<td>30K (30,920)</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>(Kotzias et al., 2015)</td>
<td>6K (6,000)</td>
</tr>
<tr>
<td>Combined</td>
<td>Diverse NLI Collection (DNC)</td>
<td>575K (576,438)</td>
</tr>
<tr>
<td></td>
<td>SNLI (Bowman et al., 2015)</td>
<td>570K</td>
</tr>
<tr>
<td></td>
<td>Multi-NLI (Williams et al., 2017)</td>
<td>433K</td>
</tr>
</tbody>
</table>
Outline

● Introduction

● The DNC: Diverse NLI Collection

● Constructing the DNC

● Experiments & Results
The DNC

<table>
<thead>
<tr>
<th>Sem. Phenomena/Annotations</th>
<th>Dataset</th>
<th># pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event Factuality</td>
<td>Decomp (Rudinger et al., 2018b)</td>
<td>42K (41,888)</td>
</tr>
<tr>
<td></td>
<td>UW (Lee et al., 2015)</td>
<td>5K (5,094)</td>
</tr>
<tr>
<td></td>
<td>MeanTime (Minard et al., 2016)</td>
<td>.7K (738)</td>
</tr>
<tr>
<td>Named Entity Recognition</td>
<td>Groningen (Bos et al., 2017)</td>
<td>260K (261,406)</td>
</tr>
<tr>
<td></td>
<td>CoNLL (Tjong Kim Sang and De Meulder, 2003)</td>
<td>60K (59,970)</td>
</tr>
<tr>
<td>Gendered Anaphora</td>
<td>Winogender (Rudinger et al., 2018a)</td>
<td>.4K (464)</td>
</tr>
<tr>
<td>Lexicosyntactic Inference</td>
<td>VerbCorner (Hartshorne et al., 2013)</td>
<td>135K (138,648)</td>
</tr>
<tr>
<td></td>
<td>MegaVeridicality (White and Rawlins, 2018)</td>
<td>11K (11,814)</td>
</tr>
<tr>
<td></td>
<td>VerbNet (Schuler, 2005)</td>
<td>2K (1,950)</td>
</tr>
<tr>
<td>Puns</td>
<td>(Yang et al., 2015)</td>
<td>9K (9,492)</td>
</tr>
<tr>
<td></td>
<td>SemEval 2017 Task 7 (Miller et al., 2017)</td>
<td>8K (8,054)</td>
</tr>
<tr>
<td>Relationship Extraction</td>
<td>FACC1 (Gabrilovich et al., 2013)</td>
<td>30K (30,920)</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>(Kotzias et al., 2015)</td>
<td>6K (6,000)</td>
</tr>
<tr>
<td>Combined</td>
<td>Diverse NLI Collection (DNC)</td>
<td>575K (576,438)</td>
</tr>
<tr>
<td></td>
<td>SNLI (Bowman et al., 2015)</td>
<td>570K</td>
</tr>
<tr>
<td></td>
<td>Multi-NLI (Williams et al., 2017)</td>
<td>433K</td>
</tr>
</tbody>
</table>
Recasting
Recasting

“Leverage existing semantic annotations to create NLI datasets that probe different semantic phenomena”
Recasting

“Leverage existing semantic annotations to create NLI datasets that probe different semantic phenomena”
“Leverage existing semantic annotations to create NLI datasets that probe different semantic phenomena”
“Leverage existing semantic annotations to create NLI datasets that probe different semantic phenomena”
Event Factuality
Event Factuality

Create natural language template
Event Factuality

Create natural language template

Extract annotated preposition
Event Factuality

Create natural language template

Extract annotated preposition

Fill in template with preposition
Event Factuality

Create natural language template

Extract annotated preposition

Fill in template with preposition

Label example based on annotation
Event Factuality

I enjoyed studying here
Event Factuality

I enjoyed studying here

happened
Event Factuality

I enjoyed studying here

The studying happened

entailed

not-entailed
Event Factuality

I enjoyed studying here

The studying did not happen

entailed
not-entailed
Event Factuality

*I actually forgot to feed my chicken*
Event Factuality

*I actually forgot to feed my chicken*

did not happened
Event Factuality

*I actually forgot to feed my chicken*

*The feeding happened*

entailed  not-entailed
Event Factuality

*I actually forgot to feed my chicken*

*The feeding did not happen*

entailed  not-entailed
Event Factuality

It Happened  (White et. al. 2016; Rudinger et. al. 2018)

42K Examples
Event Factuality

It Happened (White et. al. 2016; Rudinger et. al. 2018)
42K Examples

UW (Lee et. al. 2015)
5K Examples
Event Factuality

It Happened  (White et. al. 2016; Rudinger et. al. 2018)
42K Examples

UW  (Lee et. al. 2015)
5K Examples

MeanTime  (Minard et. al. 2016)
700 Examples
VerbNet Thematic Roles
No Comments

floss-41.2.1
Members: 4, Frames: 4

<table>
<thead>
<tr>
<th>Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRUSH (fn 1; wn 3; g 1)</td>
</tr>
<tr>
<td>FLOSS (fn 1; wn 1)</td>
</tr>
<tr>
<td>SHAVE (fn 1; wn 2; g 1)</td>
</tr>
<tr>
<td>WASH (fn 1; wn 2, 3; g 1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Roles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent [+animate]</td>
</tr>
<tr>
<td>Patient [+body_part]</td>
</tr>
<tr>
<td>Instrument</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP V NP</td>
</tr>
<tr>
<td><strong>Example</strong></td>
</tr>
<tr>
<td><strong>Syntax</strong></td>
</tr>
<tr>
<td><strong>Semantics</strong></td>
</tr>
</tbody>
</table>

| NP V |
| **Example** | "I flossed." |
| **Syntax** | Agent V |
| **Semantics** | take_care_of(during(E), Agent, ?Patient) |
1. Align tokens to Thematic Roles

MEMBERS

<table>
<thead>
<tr>
<th>Token</th>
<th>Thematic Roles</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRUSH</td>
<td>(fn 1; wn 3; g 1)</td>
</tr>
<tr>
<td>FLOSS</td>
<td>(fn 1; wn 1)</td>
</tr>
<tr>
<td>SHAVE</td>
<td>(fn 1; wn 2; g 1)</td>
</tr>
<tr>
<td>WASH</td>
<td>(fn 1; wn 2, 3; g 1)</td>
</tr>
</tbody>
</table>

ROLES

- **Agent** [+animate]
- **Patient** [+body_part]
- **Instrument**

FRAMES

**NP V NP**

<table>
<thead>
<tr>
<th>Example</th>
<th>Syntax</th>
<th>Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;The hygienist flossed my teeth.&quot;</td>
<td>Agent V Patient</td>
<td>TAKE_CARE_OF(during(E), Agent, Patient)</td>
</tr>
</tbody>
</table>

**NP V**

<table>
<thead>
<tr>
<th>Example</th>
<th>Syntax</th>
<th>Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;I flossed.&quot;</td>
<td>Agent V</td>
<td>TAKE_CARE_OF(during(E), Agent, ?Patient)</td>
</tr>
</tbody>
</table>
Recast VerbNet

1. Align tokens to Thematic Roles

**MEMBERS**

- **BRUSH** (FN 1; WN 3; G 1)
- **FLOSS** (FN 1; WN 1)
- **SHAVE** (FN 1; WN 2; G 1)
- **WASH** (FN 1; WN 2, 3; G 1)

**ROLES**

- Agent [+animate]
- Patient [+body_part]
- Instrument

**FRAMES**

**NP V NP**

**EXAMPLE** "The hygienist flossed my teeth."

**SYNTAX** Agent V Patient

**SEMANTICS** take_care_of(during(E), Agent, Patient)

**NP V**

**EXAMPLE** "I flossed."

**SYNTAX** Agent V

**SEMANTICS** take_care_of(during(E), Agent, Patient)
<table>
<thead>
<tr>
<th>Members</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BRUSH (FN 1; WN 3; G 1)</td>
<td></td>
</tr>
<tr>
<td>FLOSS (FN 1; WN 1)</td>
<td></td>
</tr>
<tr>
<td>SHAVE (FN 1; WN 2; G 1)</td>
<td></td>
</tr>
<tr>
<td>WASH (FN 1; WN 2, 3; G 1)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Roles</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AGENT [+ANIMATE]</td>
<td></td>
</tr>
<tr>
<td>PATIENT [+BODY_PART]</td>
<td></td>
</tr>
<tr>
<td>INSTRUMENT</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frames</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NP V NP</td>
<td></td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>&quot;The hygienist flossed my teeth.&quot;</td>
</tr>
<tr>
<td>SYNTAX</td>
<td>AGENT V PATIENT</td>
</tr>
<tr>
<td>SEMANTICS</td>
<td>TAKE_CARE_OF(DURING(E), AGENT, PATIENT)</td>
</tr>
<tr>
<td>NP V</td>
<td></td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>&quot;I flossed.&quot;</td>
</tr>
<tr>
<td>SYNTAX</td>
<td>AGENT V</td>
</tr>
<tr>
<td>SEMANTICS</td>
<td>TAKE_CARE_OF(DURING(E), AGENT, ?PATIENT)</td>
</tr>
</tbody>
</table>
1. Align tokens to Thematic Roles

<table>
<thead>
<tr>
<th>Agent</th>
<th>Patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Took care of</td>
<td>Patient</td>
</tr>
</tbody>
</table>

2. Convert semantics into natural language templates

**NP V NP**

<table>
<thead>
<tr>
<th>EXAMPLE</th>
<th>SEMANTICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;The hygienist flossed my teeth.&quot;</td>
<td><code>take_care_of(during(E), Agent, Patient)</code></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EXAMPLE</th>
<th>SEMANTICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;I flossed.&quot;</td>
<td><code>take_care_of(during(E), Agent, Patient)</code></td>
</tr>
</tbody>
</table>
1. Align tokens to Thematic Roles

2. Convert semantics into natural language templates

3. Fill in natural language templates
VerbNet Thematic Roles

The hygienist flossed my teeth

Agent took care of Patient

entailed    not-entailed
VerbNet Thematic Roles

The hygienist flossed my teeth

entailed
not-entailed
VerbNet Thematic Roles

*The hygienist flossed my teeth*

*Patient took care of Agent*

entailed  not-entailed
VerbNet Thematic Roles

*The hygienist flossed my teeth*

*My teeth took care of the hygienist*

entailed           not-entailed
Outline

● Introduction
● The DNC: Diverse NLI Collection
● Constructing the DNC
● Experiments & Results
Experimental Goal
Experimental Goal

“demonstrate how the DNC can help to evaluate how well models capture different types of semantic reasoning necessary for general language understanding”
Typical NLI Model

- Sentence encoder over context sentence
- Sentence encoder over hypothesis sentence
- Fully connected layer
- \( n \)-way softmax
InferSent (Conneau et al. 2017)

- Sentence encoder over context sentence
- Sentence encoder over hypothesis sentence

- Fully connected layer
- N-way softmax

u_v
A diagram showing a machine learning model architecture. The model consists of:

- GloVe embeddings
- Sentence encoder over context sentence
- Sentence encoder over hypothesis sentence
- Fully connected layer
- n-way softmax
$n$-way softmax

fully connected layer

$u, v$

Bidirectional LSTM

sentence encoder over context sentence

sentence encoder over hypothesis sentence

GloVe embeddings
Bidirectional LSTM

Max Pooling

u, v

fully connected layer

n-way softmax

sentence encoder over context sentence

sentence encoder over hypothesis sentence

GloVe embeddings
GloVe embeddings

Bidirectional LSTM

Max Pooling

MLP with 1 hidden layer

n-way softmax

fully connected layer

sentence encoder over context sentence

sentence encoder over hypothesis sentence

GloVe embeddings
Hypothesis
Only baseline (Poliak et. al. *SEM 2018)

sentence encoder over hypothesis sentence

v

fully connected layer

n-way softmax
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Recast Data</th>
<th>NER</th>
<th>EF</th>
<th>RE</th>
<th>Puns</th>
<th>Sentiment</th>
<th>GAR</th>
<th>VC</th>
<th>MV</th>
<th>VN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority (MAJ)</td>
<td></td>
<td>50.00</td>
<td>50.00</td>
<td>59.53</td>
<td>50.00</td>
<td>50.00</td>
<td>50.00</td>
<td>50.00</td>
<td>66.67</td>
<td>53.66</td>
</tr>
<tr>
<td></td>
<td>No Pre-training</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InferSent</td>
<td></td>
<td>92.50</td>
<td>83.07</td>
<td>61.89</td>
<td>60.36</td>
<td>50.00</td>
<td></td>
<td>88.60</td>
<td>85.96</td>
<td>46.34</td>
</tr>
<tr>
<td>Hyp-only</td>
<td></td>
<td>91.48</td>
<td>69.14</td>
<td>64.78</td>
<td>60.36</td>
<td>50.00</td>
<td></td>
<td>76.82</td>
<td>77.83</td>
<td>46.34</td>
</tr>
<tr>
<td></td>
<td>Pre-trained DNC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InferSent (update)</td>
<td></td>
<td>92.47</td>
<td>83.86</td>
<td>74.38</td>
<td>93.17</td>
<td>81.00</td>
<td></td>
<td>89.00</td>
<td>85.62</td>
<td>76.83</td>
</tr>
<tr>
<td>InferSent (fixed)</td>
<td></td>
<td>92.20</td>
<td>81.07</td>
<td>74.11</td>
<td>87.76</td>
<td>77.33</td>
<td></td>
<td>88.59</td>
<td>83.84</td>
<td>67.68</td>
</tr>
<tr>
<td>Hyp-only (update)</td>
<td></td>
<td>91.60</td>
<td>71.07</td>
<td>70.57</td>
<td>60.02</td>
<td>46.83</td>
<td></td>
<td>76.78</td>
<td>77.83</td>
<td>68.90</td>
</tr>
<tr>
<td>Hyp-only (fixed)</td>
<td></td>
<td>91.37</td>
<td>69.74</td>
<td>65.97</td>
<td>56.44</td>
<td>48.17</td>
<td></td>
<td>76.78</td>
<td>77.83</td>
<td>59.15</td>
</tr>
<tr>
<td></td>
<td>Pre-trained Multi-NLI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InferSent (update)</td>
<td></td>
<td>92.37</td>
<td>83.03</td>
<td>76.08</td>
<td>92.48</td>
<td>83.50</td>
<td></td>
<td>88.45</td>
<td>85.11</td>
<td>78.05</td>
</tr>
<tr>
<td>InferSent (fixed)</td>
<td></td>
<td>52.99</td>
<td>54.88</td>
<td>66.75</td>
<td>56.04</td>
<td>56.50</td>
<td></td>
<td>45.33</td>
<td>55.92</td>
<td>45.73</td>
</tr>
<tr>
<td>Hyp-only (update)</td>
<td></td>
<td>91.62</td>
<td>70.64</td>
<td>69.91</td>
<td>60.36</td>
<td>49.33</td>
<td></td>
<td>76.82</td>
<td>77.83</td>
<td>68.29</td>
</tr>
<tr>
<td>Hyp-only (fixed)</td>
<td></td>
<td>52.55</td>
<td>66.33</td>
<td>52.96</td>
<td>60.59</td>
<td>50.00</td>
<td></td>
<td>41.31</td>
<td>46.28</td>
<td>48.78</td>
</tr>
</tbody>
</table>
Experimental Setup
Experimental Setup

Train models on each DNC dataset
Experimental Setup

Train models on each DNC dataset

Pre-train models on all of DNC or Multi-NLI
Experimental Setup

Train models on each DNC dataset

Pre-train models on all of DNC or Multi-NLI

Evaluate fixed models trained on all of DNC or Multi-NLI
Summary
Summary

The *DNC: Diverse NLI Collection*
Summary

The **DNC: Diverse NLI Collection**

Convert 13 existing datasets into NLI covering 7 semantic phenomena
Summary

The **DNC: Diverse NLI Collection**

Convert 13 existing datasets into NLI covering 7 semantic phenomena

Over half a million examples
Summary

The **DNC: Diverse NLI Collection**

Convert 13 existing datasets into NLI covering 7 semantic phenomena

Over half a million examples

Presented use case of DNC
Call to the Community
Call to the Community

Dataset creators:
Call to the Community

Dataset creators:
convert your data into NLI
Call to the Community

Dataset creators:
convert your data into NLI
included in future DNC releases
Call to the Community

Dataset creators:
convert your data into NLI
included in future DNC releases

Model creators:
Call to the Community

Dataset creators:
convert your data into NLI
included in future DNC releases

Model creators:
test your models ability to capture
diverse types of reasoning
On the Evaluation of Semantic Phenomena in NMT Using NLI

(Poliak et. al. NAACL 2018)
Diverse Natural Language Inference Collection - NLI dataset that can be used to evaluate how well models perform distinct types of reasoning (EMNLP 2018)  

http://decomp.io/projects/diverse-nat...

natural-language-processing  natural-language-inference  computational-semantics  emnlp2018

6 commits  1 branch  1 release  1 contributor

Branch: master  New pull request

azpoliax  update README.md - inference is everything data

- dev
- test
- train
- README
- additional_readmes.md
- inference_is_everything.zip

DNC: Diverse Natural Language Inference Collection

Dataset associated and released as part of Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation (EMNLP 2018).
Diverse Natural Language Inference Collection - NLI dataset that can be used to evaluate how well models perform distinct types of reasoning (EMNLP 2018)  [http://decomp.io/projects/diverse-nat...](http://decomp.io/projects/diverse-nat...)

natural-language-processing  natural-language-inference  computational-semantics  emnlp2018

6 commits  1 branch  1 release  1 contributor

Branch: master  New pull request

azpoliak update README.md - inference is everything data

- dev  Released DNC and updated README  2 months ago
- test  Released DNC and updated README  2 months ago
- train  Released DNC and updated README  2 months ago
- README.md  update README.md - inference is everything data  a month ago
- additional_references.md  added bibs for original datasets  2 months ago
- inference_is_everything.zip  included White et al's IJCNLP 2017 recast data  a month ago

DNC: *Diverse Natural Language Inference Collection*

Dataset associated and released as part of *Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation* (EMNLP 2018).
Data Example

{
  "binary-label": false,
  "context": "The hygienist flossed my teeth.",
  "hypothesis": "My teeth took care of the hygienist.",
  "label": "not-entailed",
  "label-set": [
    "entailed",
    "not-entailed"
  ],
  "pair-id": 504820,
  "split": "dev",
  "type-of-inference": "Thematic Roles"
},

MetaData Example

{
    "corpus": "VerbNet",
    "corpus-license": "http://verbs.colorado.edu/verbnet",
    "corpus-sent-id": "floss-41.2.1_NP V NP",
    "creation-approach": "automatic",
    "misc": {
        "descriptionNumber": "0.2",
        "secondary": "Transitive",
        "xtag": ""
    },
    "pair-id": 504820
}
Structure of json files:

Data files:

Each datafile has the following keys and values:

- **context**: The context sentence for the NLI pair. The context is already tokenized.
- **hypothesis**: The hypothesis sentence for the NLI pair. The hypothesis is already tokenized.
- **label**: The label for the NLI pair
- **label-set**: The set of possible labels for the specific NLI pair
- **binary-label**: A True or False label. See the paper for details on how we convert the label into a binary label.
- **split**: This can be train, dev, or test.
- **type-of-inference**: A string indicating what type of inference is tested in this example.
- **pair-id**: A unique integer id for the NLI pair. The pair-id is used to find the corresponding metadata for any given NLI pair

_metadata files:

- **pair-id**: A unique integer id for the NLI pair.
- **corpus**: The original corpus where this example came from.
- **corpus-sent-id**: The id of the sentence (or example) in the original dataset that we recast.
- **corpus-license**: The license for the data from the original dataset.
- **creation-approach**: Determines the method used to recast this example. Options are automatic, manual, or human-labeled.
- **misc**: A dictionary of other relevant information. This is an optional field.
Thank you!
Data and paper available
decomp.io