Infusing Finetuning with Semantic Dependencies

Zhaofeng Wu, Hao Peng, and Noah Smith

TACL 2021; 9:226–242
It was not bad.
Motivation

It was not bad.

Pretrained Transformer

Devlin et al. (2018); Liu et al. (2019b)
Motivation

It was not bad.

Hewitt and Manning (2019); Tenney et al. (2019); Liu et al. (2019a)
Would semantics help?
Introduction
Introduction

- We show BERT/RoBERTa less prominently surface semantics...
Introduction

- We show BERT/RoBERTa less prominently surface semantics…
- … and the explicit incorporation of semantic information:
Introduction

• We show BERT/RoBERTa less prominently surface semantics…
• … and the explicit incorporation of semantic information:
  1. Improves downstream task performance
Introduction

- We show BERT/RoBERTa less prominently surface semantics…
- … and the explicit incorporation of semantic information:
  1. Improves downstream task performance
  2. Helps guard against frequent yet invalid heuristics
Introduction

• We show BERT/RoBERTa less prominently surface semantics…
• … and the explicit incorporation of semantic information:
  1. Improves downstream task performance
  2. Helps guard against frequent yet invalid heuristics
  3. Better captures nuanced linguistic phenomena
Introduction

• We show BERT/RoBERTa less prominently surface semantics…

• … and the explicit incorporation of semantic information:

  1. Improves downstream task performance
  2. Helps guard against frequent yet invalid heuristics
  3. Better captures nuanced linguistic phenomena
  4. Increases training sample efficiency
Operationalizing “Meaning”

This technique is impossible to adopt.

DELPH-IN MRS-Derived Dependencies ([DM]; Ivanova et al., 2012)
Operationalizing “Meaning”

This technique is impossible to adopt.

DELPH-IN MRS-Derived Dependencies (DM; Ivanova et al., 2012)

This technique is impossible to adopt.

Stanford Dependencies (SD; de Marneffe et al., 2006)
Probing RoBERTa with Semantics

This technique is impossible to adopt.

Probing model (Shi et al., 2016; Adi et al., 2017)
Probing RoBERTa with Semantics

This technique is impossible to adopt.

Probing model (Shi et al., 2016; Adi et al., 2017)

Ceiling model (Dozat and Manning, 2017, 2018)
Probing RoBERTa with Semantics

![Graph showing performance metrics for Probing - Ceiling; RoBERTa-base]

- **Absolute Δ:**
  - SD (syntactic): -13.5
  - DM (semantic): -23.5

- **Relative Δ (%):**
  - SD (syntactic): -24.9
  - DM (semantic): -24.9

Legend:
- SD (syntactic)
- DM (semantic)
Can we use semantics to augment pretrained transformers?
Semantics-Infused Finetuning (SIFT)

I so love to make slides.
Semantics-Infused Finetuning (SIFT)

Max Pooling

RoBERTa

I so love to make slides.
I so love to make slides.

Semantics-Infused Finetuning (SIFT)
Semantics-Infused Finetuning (SIFT)

I so love to make slides.
Semantics-Infused Finetuning (SIFT)

I so love to make slides.
Semantics-Infused Finetuning (SIFT)

I so love to make slides.
I so love to make slides.

Semantics-Infused Finetuning (SIFT)
Experiments

- Dataset: GLUE (Wang et al., 2018)
- Backbone: RoBERTa (Liu et al., 2019b)
- Parser: SOTA DM parser with 92.5 labeled F1 (Che et al., 2019)
- Graph Encoder: RGCN (Schlichtkrull et al., 2017)
  - 2 layers
  - Hidden dimension $\in \{256, 512, 768\}$
- Epochs $\in \{3, 10, 20\}$, learning rate $\in \{1 \times 10^{-4}, 2 \times 10^{-5}\}$
Results
GLUE; Improvement Over RoBERTa-base

- CoLA: 1.7
- MPRC: 0.4
- RTE: 2.0
- SST-2: 0.5
- STS-B: 0.3
- QNLI: 0.2
- QQP: 0.1
- MNLI-ID: 0.2
- MNLI-OOD: 0.4
- Avg.: 0.6

SIFT and Syntax
GLUE; Improvement Over RoBERTa-base

<table>
<thead>
<tr>
<th>Task</th>
<th>SIFT</th>
<th>Syntax</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoLA</td>
<td>1.7</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>MPRC</td>
<td>0.4</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>RTE</td>
<td>2.0</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>SST-2</td>
<td>1.9</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>STS-B</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>QNLI</td>
<td>0.5</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>QQP</td>
<td>-0.2</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>MNLI-ID</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>MNLI-OOD</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.3</td>
<td>0.6</td>
<td>0.6</td>
</tr>
</tbody>
</table>
Analysis: When Do Semantic Structures Help?

• Two datasets
  
  • HANS tests if a model uses invalid reasoning heuristics (McCoy et al., 2019)
  
  • GLUE diagnostics tests the model capability in various linguistic phenomena (Wang et al., 2018)
  
• Examine a model trained on existing NLI datasets with synthetic NLI examples
Analysis: HANS Lexical Overlap

The actor stopped the banker. does not entail The banker stopped the actor.

<table>
<thead>
<tr>
<th>RoBERTa</th>
<th>SIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>68.1</td>
<td>71.0 (+2.9)</td>
</tr>
</tbody>
</table>
The judges heard the actor resigned. does not entail The judges heard the actor.

Analysis: HANS Subsequence

<table>
<thead>
<tr>
<th>RoBERTa</th>
<th>SIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.8</td>
<td>29.5 (+3.7)</td>
</tr>
</tbody>
</table>
If the actor slept, the senator ran. does not entail The actor slept.

Analysis: HANS Constituent

<table>
<thead>
<tr>
<th>RoBERTa</th>
<th>SIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>37.9</td>
<td>37.6 (-0.3)</td>
</tr>
</tbody>
</table>
If the actor slept, the senator ran.  does not entail  The actor slept.

Before the actor slept, the senator ran.  entails

<table>
<thead>
<tr>
<th>RoBERTa</th>
<th>SIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>37.9</td>
<td>37.6 (-0.3)</td>
</tr>
</tbody>
</table>
Analysis: GLUE Diagnostics

Pred-Arg Structure

**I opened the door.** entails **The door opened.**

I opened.  

I have no **pet puppy.** entails I have no **corgi pet puppy.**

I have no pet.  

SIFT scores:

<table>
<thead>
<tr>
<th></th>
<th>RoBERTa</th>
<th>SIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>I opened the door. <strong>entails The door opened.</strong></td>
<td>43.5</td>
<td>44.6 (+1.1)</td>
</tr>
<tr>
<td>I opened. does not entail</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I have no pet puppy. <strong>entails I have no corgi pet puppy.</strong></td>
<td>36.2</td>
<td>38.3 (+2.1)</td>
</tr>
<tr>
<td>I have no pet. does not entail</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Analysis: GLUE Diagnostics

<table>
<thead>
<tr>
<th></th>
<th>RoBERTa</th>
<th>SIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical Semantics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I have a dog. entails I have an animal.</td>
<td>45.6</td>
<td>44.8 (-0.8)</td>
</tr>
<tr>
<td>does not entail I have a cat.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Knowledge</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I live in Seattle. entails I live in the U.S.</td>
<td>28.0</td>
<td>26.3 (-1.7)</td>
</tr>
<tr>
<td>does not entail I live in Antarctica.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analysis: Sample Efficiency

- Use the same downsamped MNLI training set to train RoBERTa & SIFT
Analysis: Sample Efficiency

- Use the same downsampled MNLI training set to train RoBERTa & SIFT

Absolute Δ (SIFT - RoBERTa) on MNLI

- ID.:
  - 100% (392k): 0.2
  - 0.5% (1963): 1.5
  - 0.2% (785): 2.5
  - 0.1% (392): 2.6

- OOD.:
  - 100% (392k): 0.4
  - 0.5% (1963): 1.1
  - 0.2% (785): 1.8
  - 0.1% (392): 3.3
Summary

Probing - Ceiling; RoBERTa-base

<table>
<thead>
<tr>
<th></th>
<th>Absolute Δ</th>
<th>Relative Δ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD (syntactic)</td>
<td>-13.5</td>
<td>-23.5</td>
</tr>
<tr>
<td>DM (semantic)</td>
<td>-24.9</td>
<td>-24.9</td>
</tr>
</tbody>
</table>


Summary

Probing - Ceiling; RoBERTa-base

<table>
<thead>
<tr>
<th>Absolute Δ</th>
<th>Relative Δ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD (syntactic)</td>
<td>DM (semantic)</td>
</tr>
<tr>
<td>-13.5</td>
<td>-23.5</td>
</tr>
<tr>
<td>-24.9</td>
<td>-24.9</td>
</tr>
</tbody>
</table>

Max Pooling

RoBERTa

I so love to make slides.
Thank you!