AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples

Dongyeop Kang, Tushar Khot, Ashish Sabharwal, Eduard Hovy

1. Summary
Motivation:
- Homogeneity of crowd-sourced dataset: (e.g., SNLI, SQUAD)
- Limited linguistic variations (e.g., negation) & annotation artifacts (Gururangan et al., 2018)
- Homogeneity in learned models failing to cover long-tail patterns or linguistic phenomenon

Contributions:
- Using large knowledge bases to capture common linguistic phenomena (e.g., WordNet)
- GAN framework to train a robust model
- Adversarial examples allow a task-specific but model-independent approach
- Effective in small/medium training data: +2.8% on SNLI (1%), +4.7% on SciTail (100%)
- Robustness to long-tail patterns: +6.1% on negation examples in SNLI

2. Creating Adversarial Examples

<table>
<thead>
<tr>
<th>Example Type</th>
<th>Knowledge Source</th>
<th>Relation in Knowledge</th>
<th>Function</th>
<th>New Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Premise and Hypothesis</td>
<td>entertain (56.5%)</td>
<td>P: The dog did <strong>not</strong> eat all of the chickens.</td>
<td>H: The dog ate all of the chickens.</td>
<td>E</td>
</tr>
<tr>
<td>entertain (92.1%)</td>
<td>P: The red box is in the blue box.</td>
<td>H: The blue box is in <strong>the red box</strong>.</td>
<td>SUBSTITUTION x with y in a sentence (s)</td>
<td></td>
</tr>
</tbody>
</table>

Examples produced by our function:
(E.g. use synonym (air, atmosphere) for WordNet)

3. Model

- Discriminator: entailment system (Parikh et al., 2016)
- Generators: data-augmenter from \( G^\text{rule} \) and \( G^\text{2s} \)
- A simple approach but already shows gains

Algorithm 1: Training procedure for AdvEntuRe
1. pretrain discriminator \( D(\theta) \) on \( X \)
2. pretrain generators \( G^\text{rule}(\phi) \) on \( X \)
3. for number of training iterations do
4. for mini-batch \( B \leftarrow X \) do
5. generate examples from \( \mathcal{G} \)
6. \( Z \leftarrow G(B; \phi) \)
7. \( X \leftarrow Z \) s.t. \( |Z| \leq \alpha|X| \)
8. optimize discriminator:
9. \( \hat{\theta} \leftarrow \arg\min L_D(Z \leftarrow X; \phi) \)
10. optimize generator:
11. \( \hat{\phi} \leftarrow \arg\min L_G(Z \leftarrow G; \phi) \)
12. Update \( \theta \leftarrow \hat{\theta}, \phi \leftarrow \hat{\phi} \)

Adversarial training to create a robust discriminator (c.f. normal GAN for robust generator)

4. Evaluation
- Dataset: SNLI (570K) (Bowman et al., 2015), SciTail (27K) (Khot et al., 2018)
- We train on small set but also that we test on the full set.
- +6.1% on nega-SNLI test (examples containing handful of negation patterns)

Ablation

Code available at https://github.com/dykang/Adventure