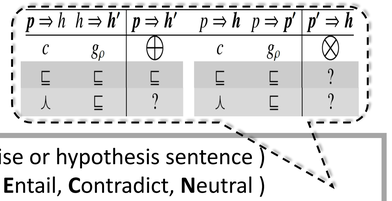
AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples

Dongyeop Kang, Tushar Khot, Ashish Sabharwal, Eduard Hovy

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Carnegie Mellon University
Language Technologies Institute



1. Summary

Motivation:

- ❑ Homogeneity of crowd-sourced dataset: (e.g., SNLI, SQUAD)
  - ❑ Limited linguistic variations (e.g., negation) & annotation artifacts (Gururangan et al., 18)
  - ❑ Homogeneity in learned models failing to cover long-tail patterns or linguistic phenomenon

Prediction (Parikh et al., 16)	Premise and Hypothesis
entails (56.5%)	**P**: The dog did not eat all of the chickens. **H**: The dog ate all of the chickens.
entails (92.1%)	**P**: The red box is in the blue box. **H**: The blue box is in the red box.

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

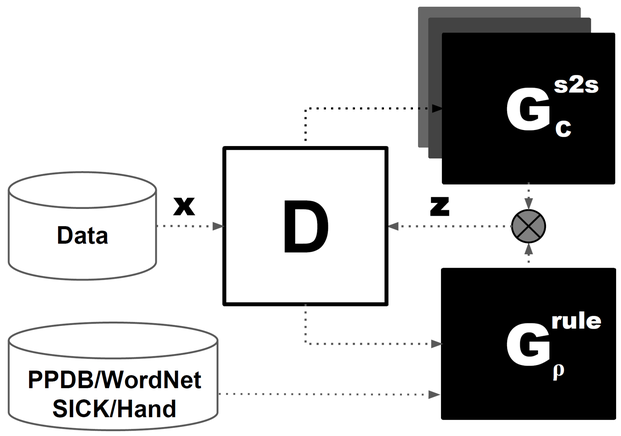
(**x**/**y** is premise or hypothesis sentence)
(**E**ntail, **C**ontradict, **N**eutral)

Example Type	Knowledge Source	Relation in Knowledge	Function	New Label
Knowledge Base	**WordNet** (Miller et al., 95)	*Hypernym (x, y)*	SUBSTITUTE x with y in a sentence (s)	*E*
		Antonym (x, y)		*C*
		Synonym (x, y)		*E*
	PPDB (Ganitkevitch et al., 13)	*x ≡ y*		*E*
	SICK (Marelli et al., 14)	***c** (x, y)*		***c***
Hand Rule	domain Knowledge	*NEGATE*	NEGATE (s)	
Neural	training Data	*(Seq2Seq, **c**)*	Seq2Seq (s)	***c***

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

Original Premise		a person on a horse jumps over a broken down airplane
Generated Hypothesis	S2S, **E**	a person is on a horse jumps over a rail, a person jumping over a plane
	S2S, **C**	a person is riding a horse in a field with a dog in a red coat
	S2S, **N**	a person is in a blue dog is in a park
Original Premise		a dirt bike rider catches some air going off a large hill
Generated Hypothesis	PPDB, **E**	a dirt **motorcycle** rider catches some air going off a large hill
	SICK, **N**	a dirt bike **man on yellow bike** catches some air going off a large hill
	$\text{Word}\text{Net}_{\text{et}}$, **E**	a dirt bike rider catches some **atmosphere** going off a large hill
	Hand, **C**	a dirt bike rider **do not catch** some air going off a large hill

3. Model



Algorithm 1 Training procedure for AdvEntuRe.

```
pretrain discriminator 𝔻(θ̂) on X;
pretrain generators 𝔾_c^{s2s}(φ̂) on X;
for number of training iterations do
  for mini-batch B ← X do
    generate examples from 𝔾
      Z_G ⇐ 𝔾(B; φ),
    balance X and Z_G s.t. |Z_G| ≤ α|X|
    optimize discriminator:
      θ̂ = argmin_θ L_𝔻(X + Z_G; θ)
    optimize generator:
      φ̂ = argmin_φ L_{𝔾^{s2s}}(Z_G; L_𝔻; φ)
    Update θ ← θ̂; φ ← φ̂
```

- ❑ **D**iscriminator: entailment system (Parikh et al., 16)
- ❑ **G**enerators: data-augmenter from $\mathbf{G}^{\text{rule}}$ and $\mathbf{G}^{\text{s2s}}$
- ❑ a simple approach but already shows gains

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

4. Evaluation

- ❑ Dataset: **SNLI** (570K) (Bowman et al., 15), **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)

SNLI	1%	10%	50%	100%
𝔻	57.68	75.03	82.77	84.52
$\mathbb{D}_{\text{retro}}$	57.04	73.45	81.18	84.14
AdvEntuRe				
⌞ 𝔻 + $\mathbb{G}^{\text{s2s}}$	58.35	75.66	82.91	**84.68**
⌞ 𝔻 + $\mathbb{G}^{\text{rule}}$	**60.45**	**77.11**	**83.51**	84.40
⌞ 𝔻 + $\mathbb{G}^{\text{rule}}$ + $\mathbb{G}^{\text{s2s}}$	59.33	76.03	83.02	83.25

+2.8% on SNLI (1%)

SciTail	1%	10%	50%	100%
𝔻	56.60	60.84	73.24	74.29
$\mathbb{D}_{\text{retro}}$	59.75	67.99	69.05	72.63
AdvEntuRe				
⌞ 𝔻 + $\mathbb{G}^{\text{s2s}}$	**65.78**	**70.77**	74.68	76.92
⌞ 𝔻 + $\mathbb{G}^{\text{rule}}$	61.74	66.53	73.99	**79.03**
⌞ 𝔻 + $\mathbb{G}^{\text{rule}}$ + $\mathbb{G}^{\text{s2s}}$	63.28	66.78	**74.77**	78.60

+4.7% on SciTail (100%)

	R/C	SNLI (5%)
𝔻 + $\mathbb{G}^{\text{rule}}$	𝔻	69.18
	+ PPDB	**72.81 (+3.6%)**
	+ SICK	71.32 (+2.1%)
	+ WordNet	71.54 (+2.3%)
	+ HAND	71.15 (+1.9%)
	+ all	71.31 (+2.1%)
𝔻 + $\mathbb{G}^{\text{s2s}}$	𝔻	69.18
	+ positive	71.21 (+2.0%)
	+ negative	71.76 (+2.6%)
	+ neutral	71.72 (+2.5%)
	+ all	**72.28 (+3.1%)**

Ablation

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