Generating Negative Commonsense Knowledge

tsafavi@umich.edu

Tara Safavi + Danai Koutra (University of Michigan)

INTRODUCTION

Commonsense knowledge bases (KBs): Store declarative statements of implicit human knowledge (e.g., pre-conditions, causes, properties) in relational triple form

- Ever-expanding KBs serve as *relational inductive biases* [Battaglia et al 2018]
- **KB completion**: Automatically augment KBs with novel statements
- Positive *and* **negative knowledge** needed for KB completion
- Negative knowledge: False or non-viable statements (different from *negation*)

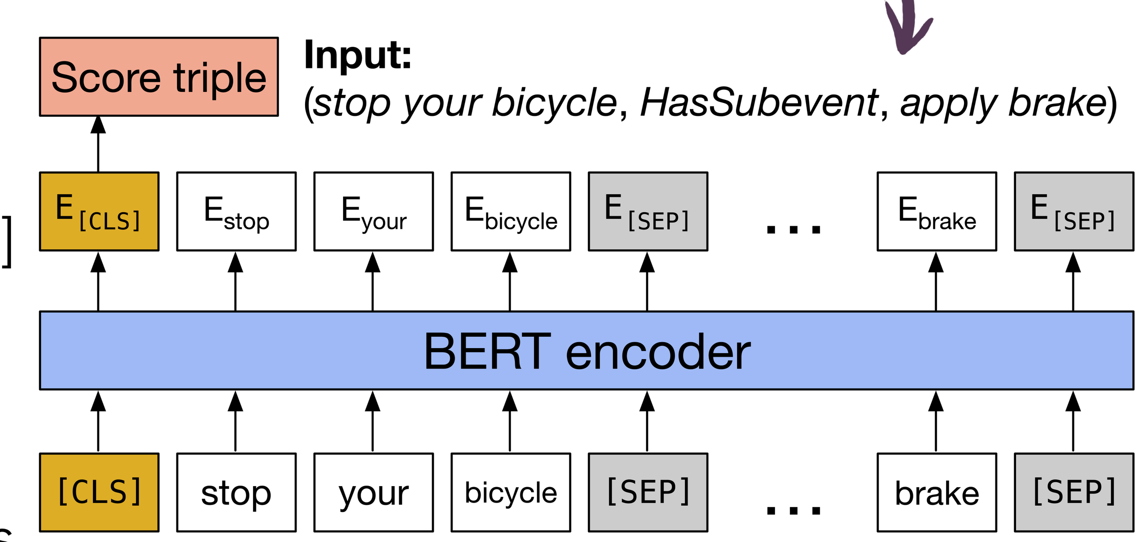
CONTRIBUTIONS

- We show the difficulty of obtaining meaningful negatives in KBs
- We propose **NegatER**, a negative knowledge generation framework
- We demonstrate the intrinsic value and extrinsic utility of negative knowledge

PRELIMINARY EXPERIMENTS

TERMINOLOGY

- Commonsense statements are KB triples: (*head phrase*, *relation*, *tail phrase*)



EXPERIMENTAL SETUP

- Classify novel triples as {True, False}
- **ConceptNet** dataset [Speer and Havasi 2012]
- *Randomly corrupted* negatives [Li et al 2016]

MODELS

- 7 self-supervised and unsupervised baselines
- We propose a **fine-tuned BERT** model with a triple scoring layer [Devlin et al 2019]

RESULTS

- BERT **beats all published results**…
- …because the **task is too easy** for BERT

~50% of test positives are paraphrases of train and ~40% of **negatives are ungrammatical**; paraphrases are easy to *delete*, but good negatives aren't easy to *construct*

EASY NEGATIVES **HARD NEGATIVES**

	Original	New	Diff.
Bilinear AVG [7]	0.9170	0.6695	-0.2475
DNN AVG [7]	0.9200	0.6410	-0.2790
DNN LSTM [7]	0.8920	0.6305	-0.2615
DNN AVG + CKBG [13]	0.9470	-	-
Factorized [6]	0.7940	0.7068	-0.0872
Prototypical [6]	0.8900	0.5586	-0.3314
Coherency Ranking [3]	0.7880	0.6387	-0.1493
BERT (ours)	**0.9537**	**0.7855**	-0.1682
Human estimate	0.95	0.86	-0.09

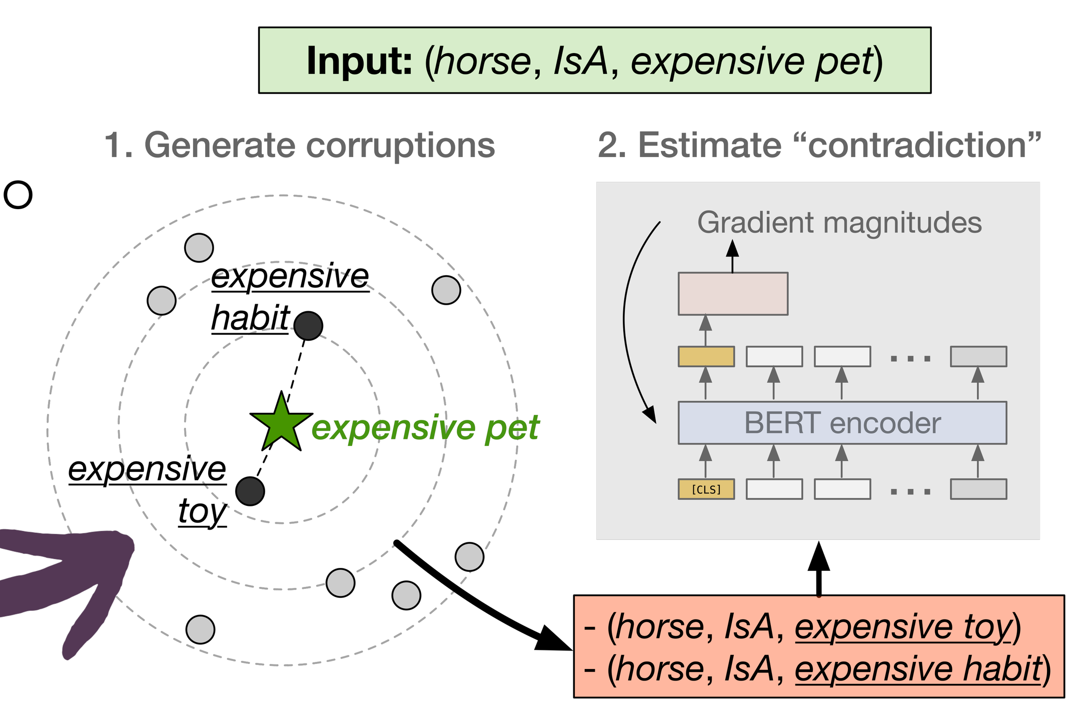
NegatER: PROPOSED FRAMEWORK

We want negatives "**on the boundary**" of positive knowledge [Minsky 1997] – knowledge that looks plausible and is "almost correct", but would be misleading or harmful if considered as true (i.e., nontrivial negatives)

STEP 1: CORRUPT POSITIVES

Given a positive triple:

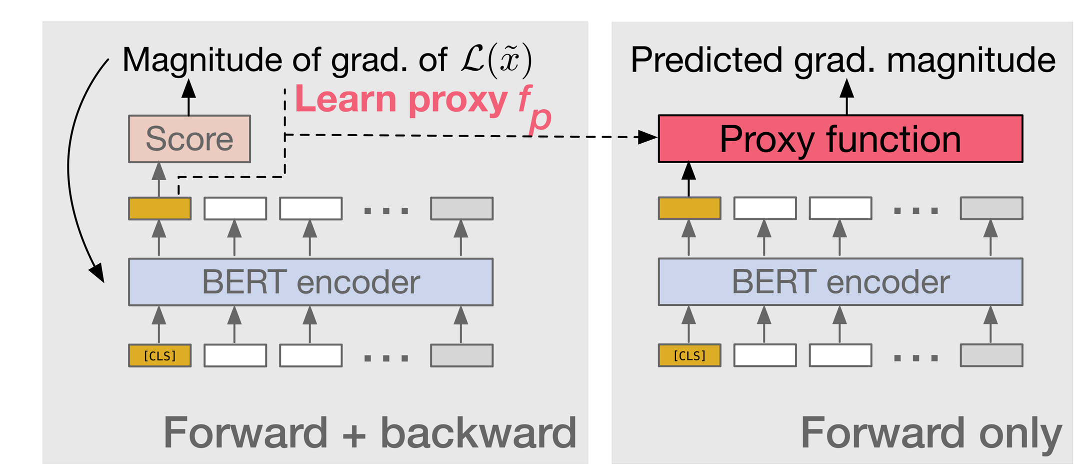
1. Retrieve top-k semantically similar phrases to head phrase (*we use pretrained BERT here*)
2. Replace head phrase with k-nearest neighbors in turn
3. Discard in-KB triples
4. Repeat for tail phrase



STEP 2: FIND CONTRADICTIONS

Rank corruptions by the **amount needed to update fine-tuned BERT's parameters given a positive labeling** (i.e., the magnitude of the gradient of the loss)

- For efficiency, learn a **proxy function to** ***predict*** **gradient magnitudes** so that backpropagation can be skipped!
- Train on triple embeddings + gradient magnitudes for a sample of corruptions



EVALUATION

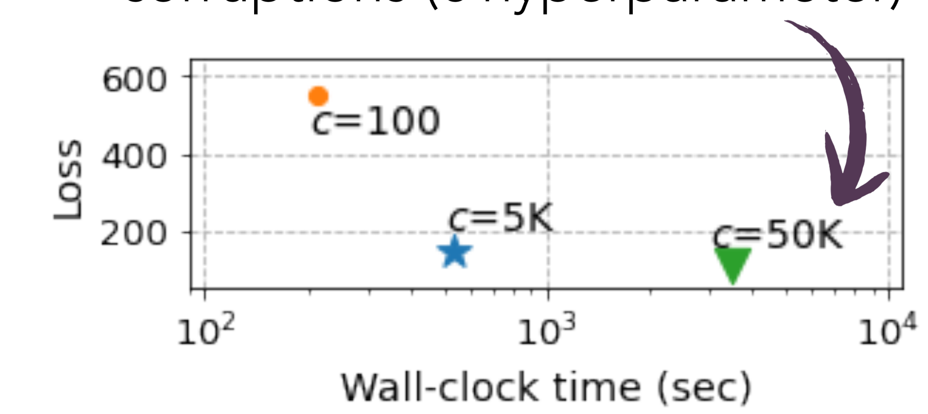
EXAMPLE GENERATED NEGATIVES

Head phrase	Relation	Tail phrase
heater	*UsedFor*	*produce breeze*
computer program	*MadeOf*	*silicon*
fly kite	*HasPrerequisite*	*get skis*
muffin	*AtLocation*	*hot-dog stand*
butterfly	*HasProperty*	*hunted by humans for food*
theatre ticket	*UsedFor*	*get home from work*

~94.5% of our negatives grammatical and ~86% true negatives, compared to 60% grammatical and 90% true negatives for random corruptions

Hard negatives significantly reduce performance of all models (-21.77 points avg), compared to human (-9 points)

Proxy approach learns good model trained on relatively few corruptions (*c* hyperparameter)



Relational Inductive Biases, Deep Learning, and Graph Networks. Peter Battaglia et al, arxiv 2018
Representing General Relational Knowledge in ConceptNet 5. Robyn Speer and Catherine Havasi, LREC 2012
Commonsense Knowledge Base Completion. Xiang Li et al, ACL 2016
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Jacob Devlin et al, NAACL 2019
Negative Expertise. Marvin Minsky, International Journal of Expert Systems 1994