Can a Gorilla Ride a Camel? Learning Semantic Plausibility from Text

Ian Porada[1], Kaheer Suleman[2], and Jackie Chi Kit Cheung[1]
[1]Mila, McGill University
{ian.porada@mail, jcheung@cs}.mcgill.ca
[2]Microsoft Research Montreal
kasulema@microsoft.com

Mila McGill Microsoft

Can a person ride a camel?

(Busson, 2007)

Three camels in a line: a person rides on the leading camel.

Mila McGill Microsoft

Can a {person}$_{subj}$ {ride}$_{verb}$ a {camel}$_{obj}$?

(Busson, 2007)

Three camels in a line:
a {person}$_{subj}$ {rides}$_{verb}$
on the leading {camel}$_{obj}$.

Mila McGill Microsoft

Can a gorilla ride a camel?

(FFNR, 2018)

???

Can a lake ride a camel?

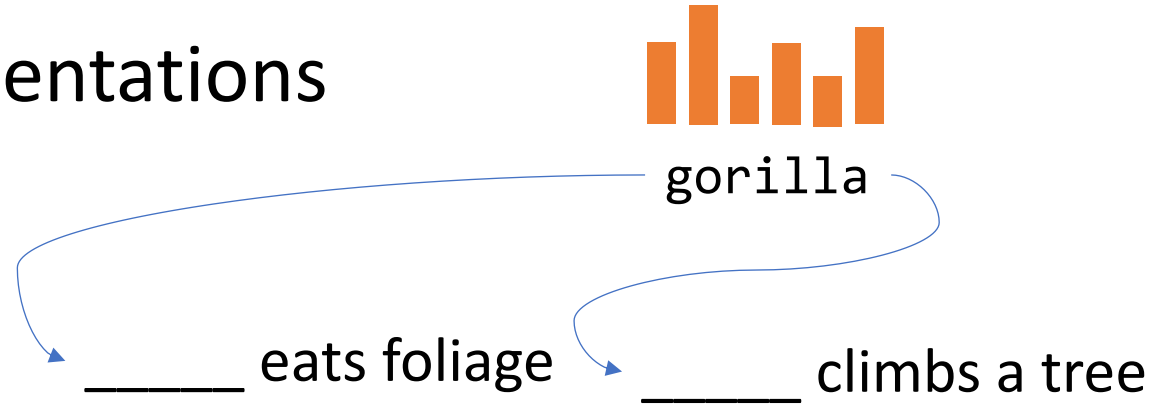
(Schmaltz, 2018)

???

Modeling semantic plausibility

- **Necessary** for many NLU tasks
  - Hard coreference resolution (Peng et al., 2015)
  - Paragraph reconstruction (Li and Jurafsky, 2017)

- A **testbed** for language representations



Existing work

- Distributional cues **fail** at modeling semantic plausibility (Wang et al., 2018)
  - We can improve performance by **injecting explicit commonsense knowledge**
    - Weight
    - Size
    - Sentience
    - …

Our points

1. Distributional representations are **sufficient** for semantic plausibility in the **supervised setting**.

2. Solving semantic plausibility without manual supervision is an interesting problem.
   - Formulation & baseline

Data

- We focus on **physical plausibility**
  - Is a given *subject-verb-object* (s-v-o) triple physically plausible?

- Wang et al.'s (2018) Physical Plausibility Dataset
  - 3,062 s-v-o triples
  - 150 verbs
  - 450 nouns

Wang et al.'s (2018) Physical Plausibility Dataset

Event	Plausible?
bird-construct-nest	✓
gorilla-ride-camel	✓
bottle-contain-elephant	✗
lake-fuse-tie	✗

Mila McGill Microsoft

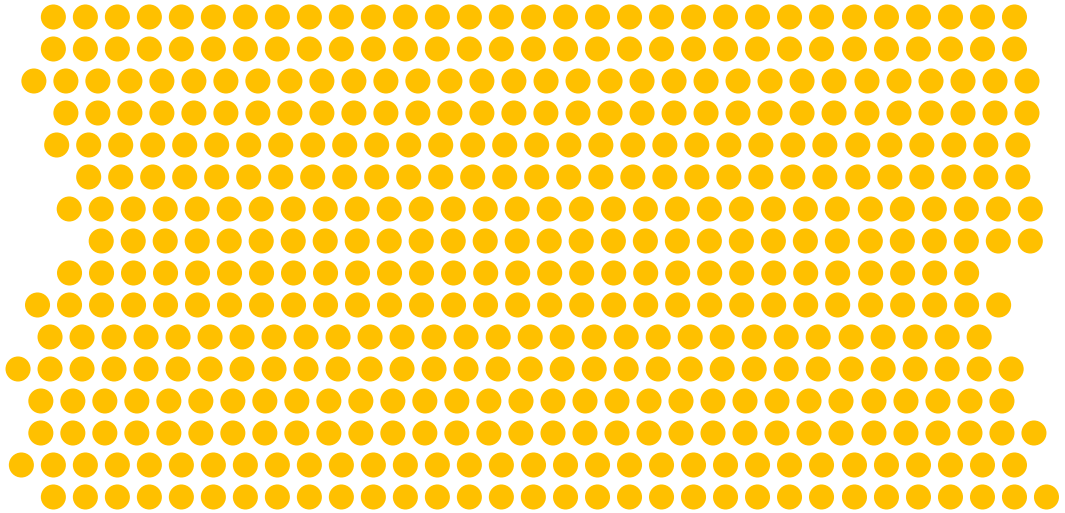
Methods

- NN (Van de Cruys, 2014)
  - Baseline
  - MLP over GloVe embeddings

- BERT (Devlin et al., 2019)
  - Large, pretrained language model
  - Treat input as a sequence, “<subj> <verb> <obj>”
  - Finetune entire model with MLP head

Supervised setting (Wang et al., 2018)

Train



Test

Results (supervised)

Model	**Accuracy**
Random	0.50
NN (Van de Cruys, 2014)	0.68
NN+WK* (Wang et al., 2018)	0.76
Fine-tuned BERT	**0.89**

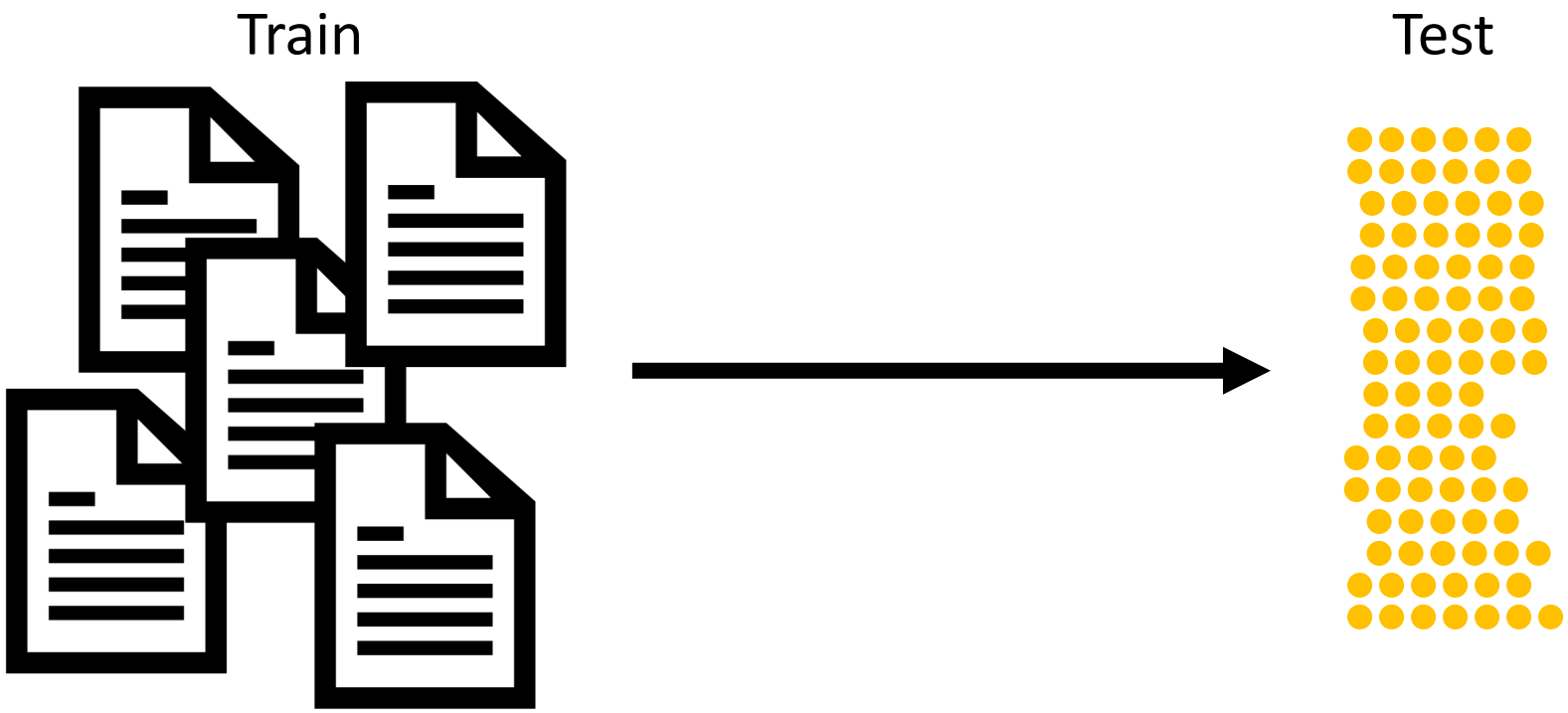
*WK = weight, size, sentience, …

Mila McGill Microsoft

Results (supervised)

- But did we solve semantic plausibility?
  - Performance depends on the **coverage** of the training set vocabulary (Moosavi and Strube, 2017)
  - Susceptible to **annotation artifacts** (Gururangan et al., 2018; Poliak et al., 2018)
  - Not necessarily learning the **desired relation** (Levy et al., 2015)

Proposed, unsupervised setting

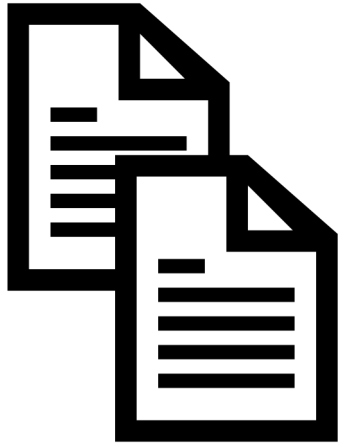


Proposed, unsupervised setting

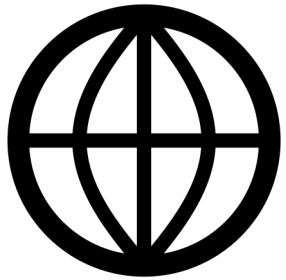
- Requires going **beyond a distributional representation**
- We take **attested** events to be **plausible**
  - *woman-ride-camel*
- And pseudo-negative **random events** to be **implausible**
  - Sample subject, verb, and object independently by occurrence frequency
  - *band-produce-camel*

Proposed, unsupervised setting

English Wikipedia



NELL 604m (Carlson et al., 2010)



Results (unsupervised)

Model	Wikipedia		NELL	
	Valid	Test	Valid	Test
Random	0.50	0.50	0.50	0.50
NN	0.53	0.52	0.50	0.51
BERT	**0.65**	**0.63**	**0.57**	**0.56**

Results (unsupervised)

- But the performance is **limited**
  - Succumbs to **reporting bias** (Gordon and Van Durme, 2013)
  - Lacks hierarchical generalization
    - *grandfather-ride-camel*
    - *teammate-ride-camel*
    - *woman-ride-camel*
    - *man-ride-camel*
  - Could be improved with better negative sampling

Conclusion

- Distributional signals **sufficient** for semantic plausibility in a **supervised setting**

- Improving performance without manual supervision is an interesting direction
  - A testbed for injecting commonsense knowledge
  - An **incidental signal** (Roth, 2017)