Modeling Event Plausibility with Consistent Conceptual Abstraction

Ian Porada[1], Kaheer Suleman[2], Adam Trischler[2], and Jackie Chi Kit Cheung[1]

[1]Mila, McGill University

{ian.porada@mail, jcheung@cs}.mcgill.ca

[2]Microsoft Research Montréal

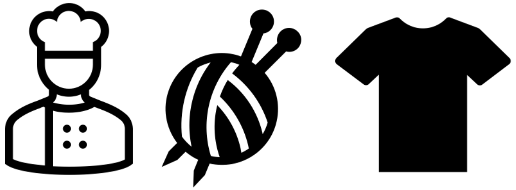
{kasulema, adam.trischler}@microsoft.com

Intro

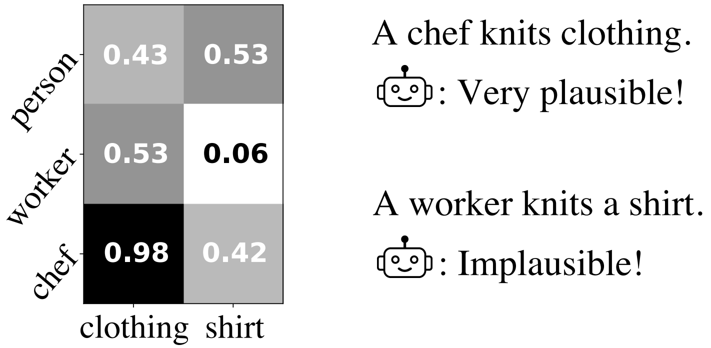
The likelihood estimates of a pre-trained language model (PTLM) are often used as a **proxy for plausibility**.

We:

- show that these estimates are **inconsistent** w.r.t. **conceptual abstractions** of an event
- explore how **enforcing consistency** can **improve correlation** with human plausibility judgements
- propose two automatic metrics of consistency



Is it plausible an [X] knits a [Y]?



Estimates of RoBERTa-base (fine-tuned to predict probability of event occurrence).

Motivation

Modeling plausibility is **implicit** in:

- **Coreference resolution** (Hobbs, 1978; Dagan and Itai, 1990; Zhang et al., 2019b)
- **Word sense disambiguation** (Resnik, 1997; McCarthy and Carroll, 2003)
- **Textual entailment** (Zanzotto et al., 2006; Pantel et al., 2007)
- **Semantic role labeling** (Gildea and Jurafsky, 2002; Zapirain et al., 2013)
- **Commonsense inference** (Gordon et al., 2011; Zhang et al., 2017; Bhagavatula et al., 2020)

Motivation

Understanding natural language requires discerning **plausible** and **implausible** events (Wilks, 1975).


E.g., The car is filled with gas, and I’m starting to breathe ***it*** in.

Breathe what in?

The gas. (Breathing in a car is implausible.)

Background

Selectional preferences (Evens, 1975; Resnik, 1993; Erk et al., 2010)

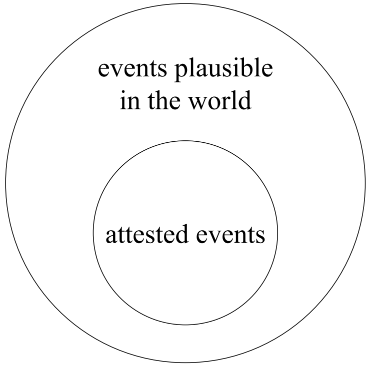
The **semantic preference** of a predicate for an argument (in a certain relation)

- E.g., the relative preference of *knit* for the direct object *shirt*

Reporting bias (Gordon and Van Durme, 2013)

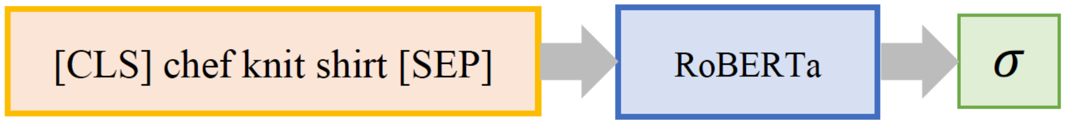
Common events are under-reported in text

- E.g., “a person breathes” is less likely to be attested in a corpus than “a person dies”

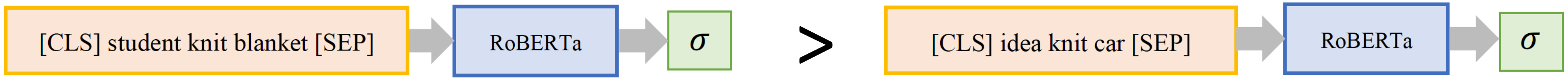


Baseline

Let an event be represented as a subject-verb-object (s-v-o) triple.

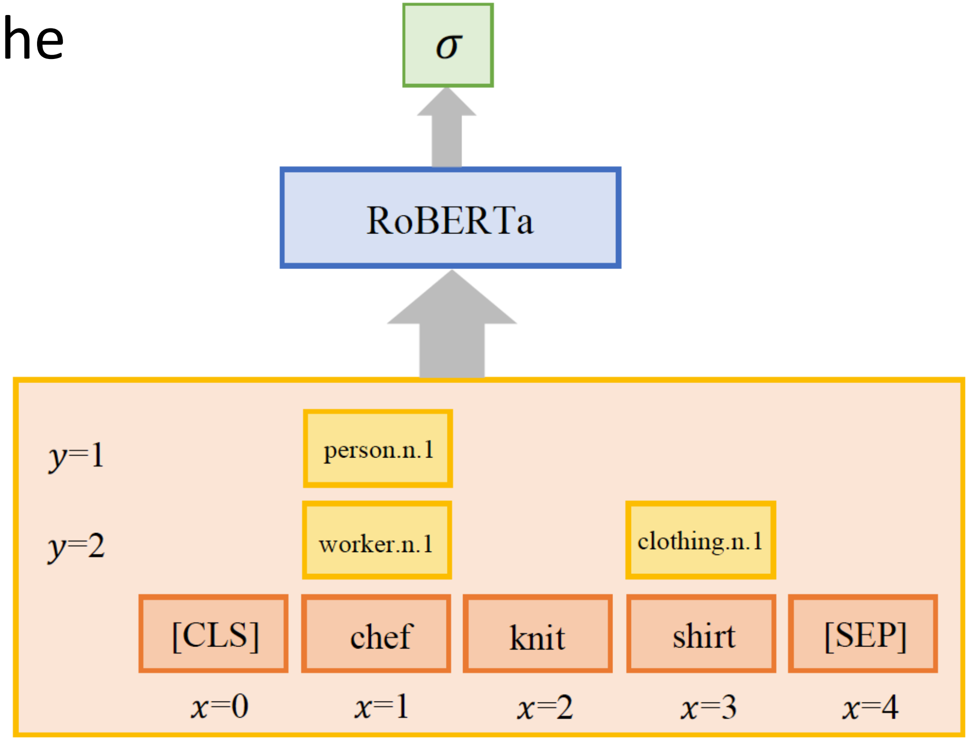


Fine-tune RoBERTa to predict **probability of occurrence** of an event in a corpus.



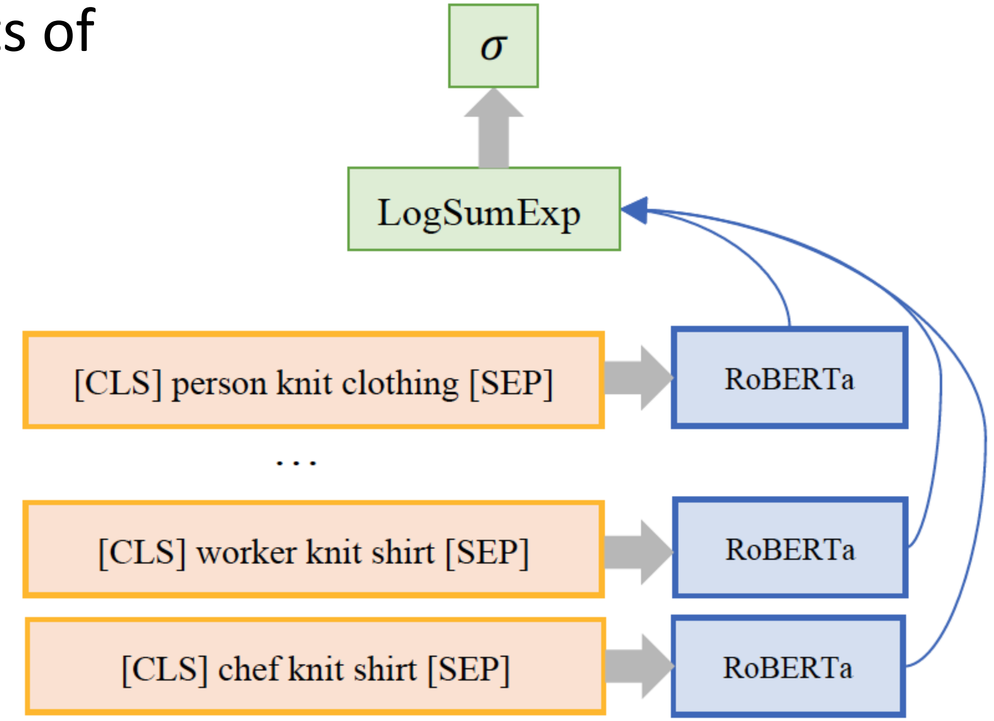
ConceptInject

- Introduce a new **token** for each hypernym
  - Initialize by average word embedding of the hypernym's lemma
- Include all hypernyms **in the input** at training and inference with a new position embedding

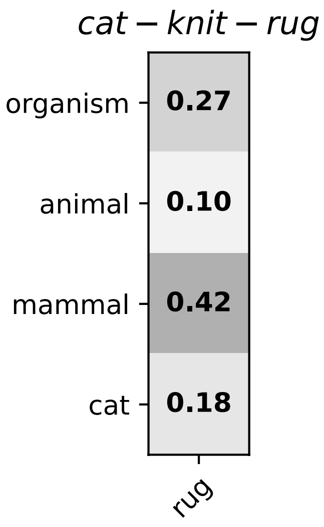


ConceptMax

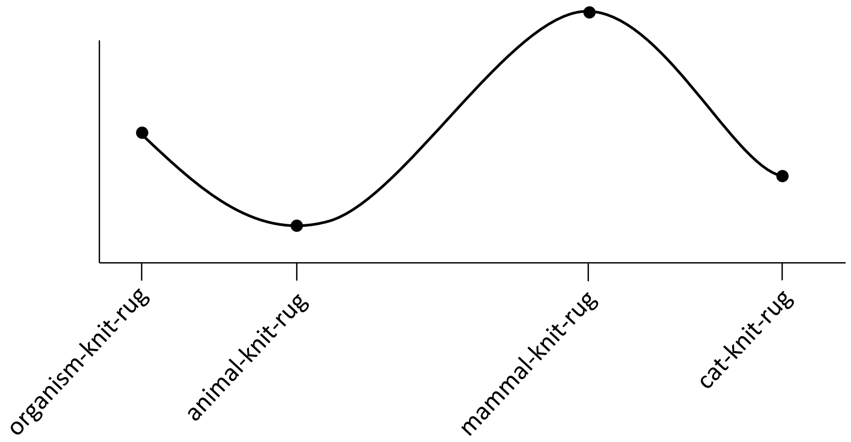
- Take the plausibility estimate of an event to be **a soft maximum** over the RoBERTa outputs of all abstractions
  - We **sample** three abstractions for tractability
- At inference, take a hard maximum over abstractions



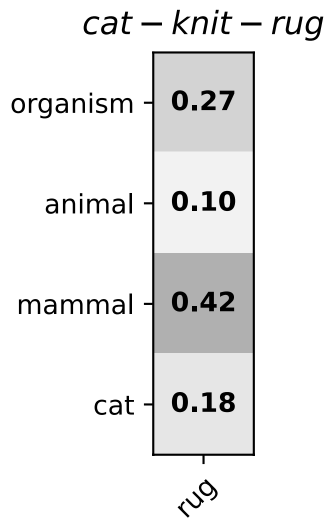
Proposed consistency metrics



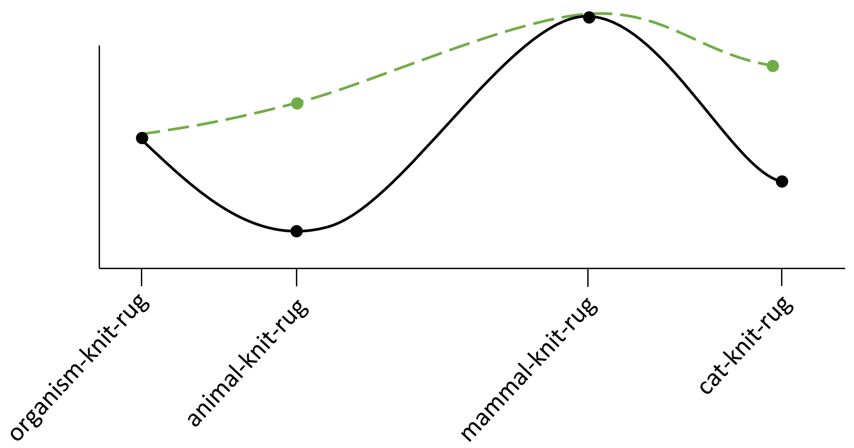
(RoBERTa estimates.)



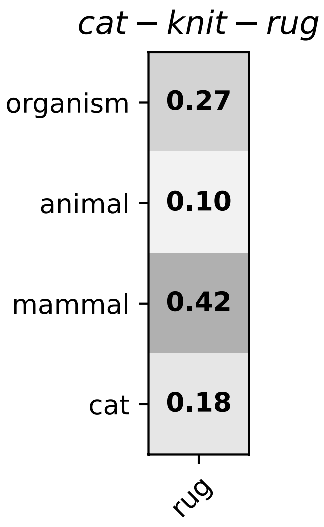
Proposed consistency metrics



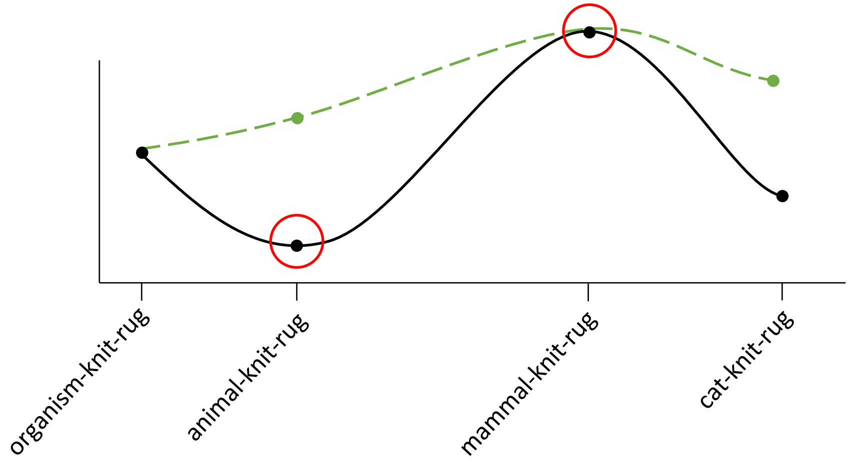
(RoBERTa estimates.)



Proposed consistency metrics

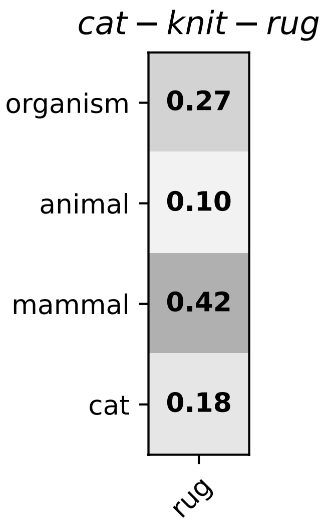


(RoBERTa estimates.)

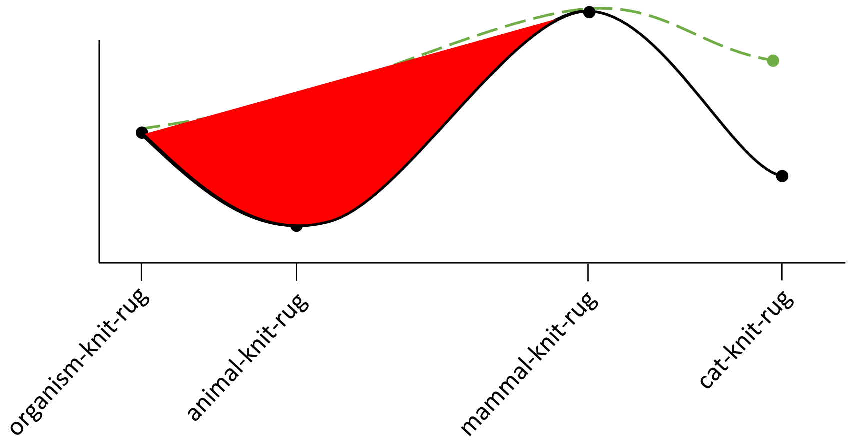


Local Extremum Rate (LER): What percentage of plausibility estimates are local extrema?

Proposed consistency metrics



(RoBERTa estimates.)



Local Extremum Rate (LER): What percentage of plausibility estimates are local extrema?

Concavity Delta (CCΔ): Are estimates for sequential abstractions (a_{i-1}, a_i, a_{i+1}) concave?

I.e., average of:

$$\delta = \begin{cases} \frac{1}{2}(a_{i-1} + a_{i+1}) - a_i & 2a_i < a_{i-1} + a_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

Training details

- Dependency parse **English Wikipedia**
    - extract s-v-o triples as training examples
        - e: attested event (positive)
        - e': random perturbation of e (negative)
- We use **WordNet 3.1** (Miller, 1995) hypernymy relations

Wikipedia
The Free Encyclopedia

e	e'
animal-eat-seed	*animal-eat-area*
passenger-ride-bus	*bus-ride-bus*
fan-throw-fruit	*group-throw-number*
woman-seek-shelter	*line-seek-issue*

Evaluation datasets

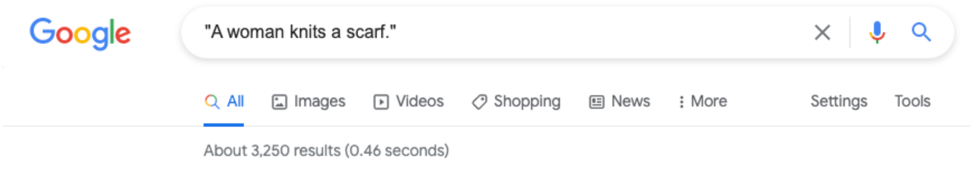
- **Physical Event Plausibility (PEP-3K)** presented by Wang et al. (2018)
- **20Q**: a subset of the Twenty Questions dataset, reformatted (github.com/allenai/twentyquestions)

PEP-3K	*chef-bake-cookie*	✓
	dog-close-door	✓
	fish-throw-elephant	✗
	marker-fuse-house	✗
20Q	*whale-breathe-air*	✓
	wolf-wear-collar	✓
	cat-hatch-egg	✗
	armrest-breathe-air	✗

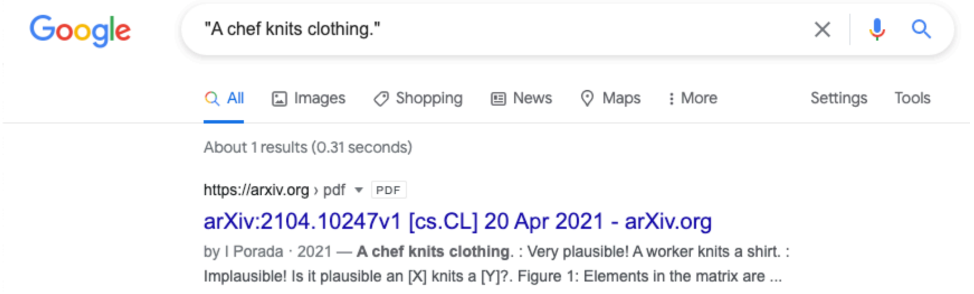
Evaluation datasets

- **Physical Event Plausibility (PEP-3K)** presented by Wang et al. (2018)
- **20Q**: a subset of the Twenty Questions dataset, reformatted (github.com/allenai/twentyquestions)

PEP-3K	*chef-bake-cookie*	✓
	dog-close-door	✓
	fish-throw-elephant	✗
	marker-fuse-house	✗
20Q	*whale-breathe-air*	✓
	wolf-wear-collar	✓
	cat-hatch-egg	✗
	armrest-breathe-air	✗



(3,250 results)



(1 result)

Results

Predicting Human Plausibility Judgements (AUC)

Model	PEP-3K	20Q	Avg.
n-gram	.51	.52	.52
GloVe+MLP	.55	.52	.53
RoBERTa$_{\text{Zero-shot}}$	.56	.57	.56
RoBERTa	.64	.67	.66
CONCEPTINJECT	.64	.66	.65
CONCEPTMAX	**.67**	**.74**	**.70**

Consistency (lower is more consistent)

	PEP-3K		20Q	
Model	CCΔ	LER	CCΔ	LER
n-gram	.06	.50	.07	.50
GloVe+MLP	.03	.61	.03	.49
RoBERTa$_{\text{Zero-shot}}$	.13	.70	.12	.65
RoBERTa	.09	.52	.08	.51
CONCEPTINJECT	.08	.52	.07	.51
CONCEPTMAX	.02	.00	.02	.00

Conclusion and future work

- Language model **estimates of occurrence** are **inconsistent** across conceptual abstractions in a lexical hierarchy
- Enforcing consistency **improves** correlation with human judgements

From here:

- How might we design a non-monotonic model of plausibility?
- How might we apply these ideas to a practical, downstream application?

Please see the paper for references and additional details.