



# Investigating Failures to Generalize for Coreference Resolution Models

Ian Porada<sup>1</sup>, Alexandra Olteanu<sup>2</sup>, Kaheer Suleman<sup>2</sup>, Adam Trischler<sup>2</sup>, and Jackie Chi Kit Cheung<sup>1</sup>

<sup>1</sup>Mila, McGill University

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

<sup>2</sup>Microsoft Research Montréal

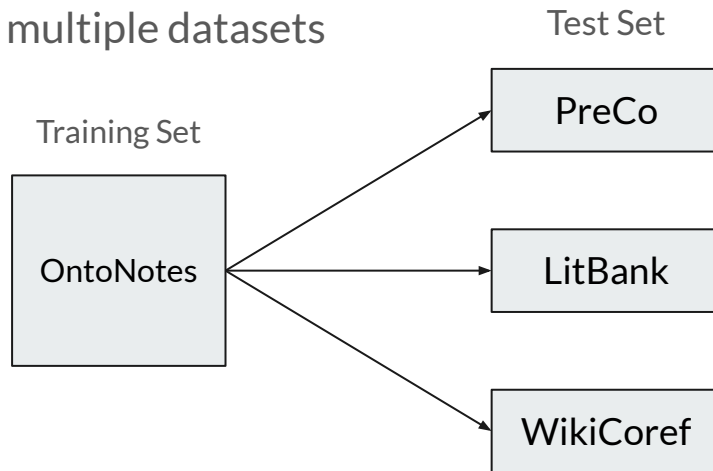
{alexandra.olteanu, kasulema, adam.trischler}@microsoft.com

# Background

Recent work evaluates **generalization** using multiple datasets

E.g., Toshniwal et al. (CRAC 2021):

(See also: Bamman et al., LREC 2020; Xia & Van Durme, EMNLP 2021; Žabokrtský et al., CRAC 2022; i.a.)





# Background

Datasets vary in how coreference is annotated; e.g.

## OntoNotes Guidelines

Coreferring generic mentions *are not* annotated

“Dogs are friendly, and dogs often bark.”

## PreCo Guidelines

Coreferring generic mentions *are* annotated

“**[Dogs]**<sub>1</sub> are friendly, and **[dogs]**<sub>1</sub> often bark.”

# Background

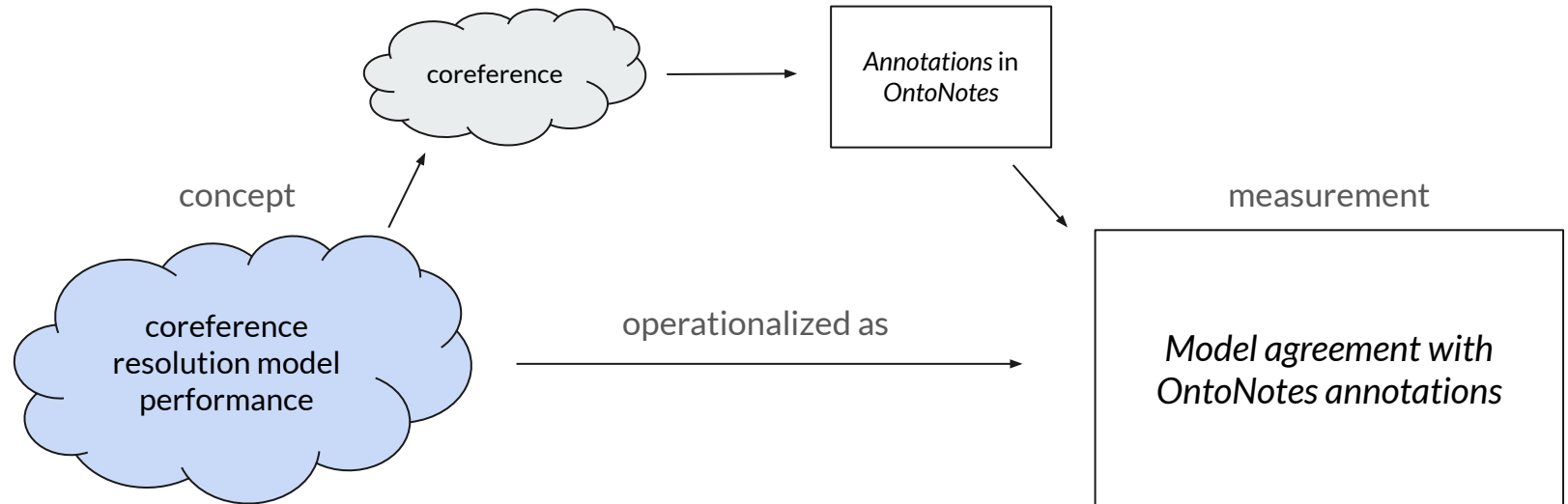
Construct validity (e.g., Adcock & Collier, APSR 2001; Jacobs & Wallach, FAccT 2021):

*Are measurements of a concept meaningful and useful?*



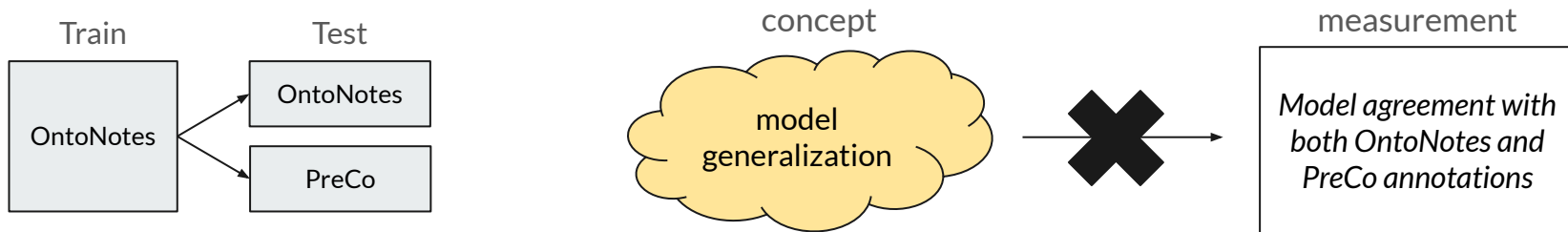
# Background

E.g., standard, in-domain evaluation:



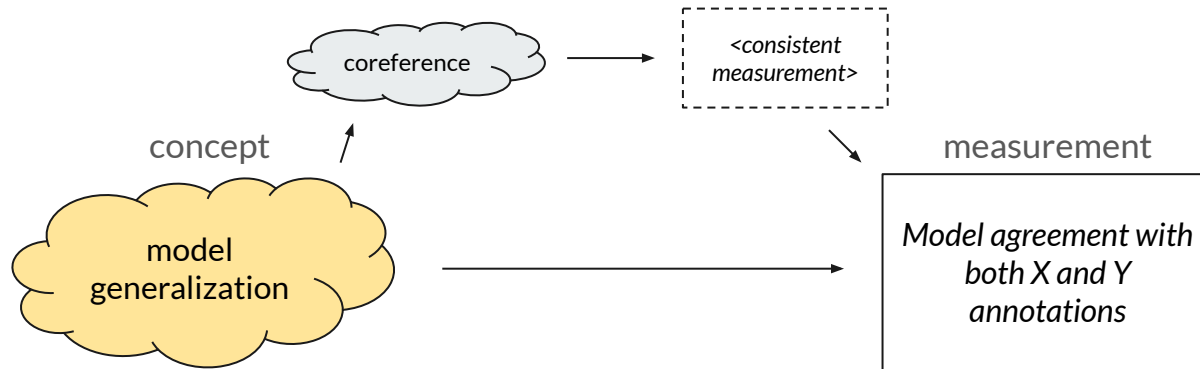
# Our Main Claim

- Measurements of **model generalization** are not accurately measuring the intended concept



# Our Main Claim

- Measurements of **model generalization** are not accurately measuring the intended concept
- Valid measurements require **resolving inconsistencies** in how coreference is **operationalized** across datasets



# Evidence

1. Models perform poorly on certain types of coreference. E.g., generic mentions:



“**[Dogs]<sub>1</sub>** are friendly, and  
**[dogs]<sub>1</sub>** often bark.”

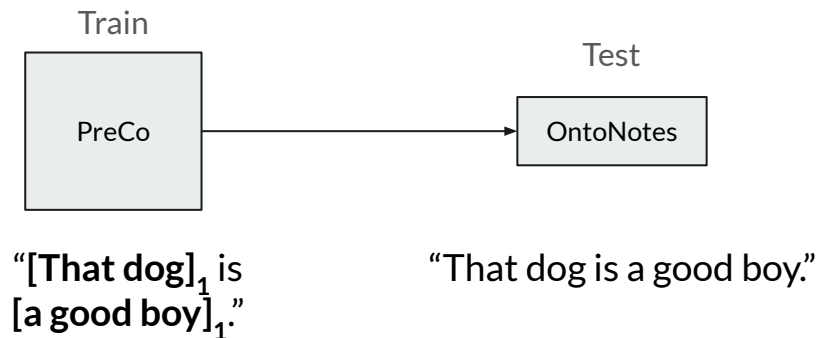
“Dogs are friendly, and  
dogs often bark.”

	PreCo F1	OntoNotes F1	$\Delta$
Overall	85.3	64.4	↓20.9 (-24%)
Generics	76.1	11.9	↓64.2 (-84%)



# Evidence

1. Models perform poorly on certain types of coreference. E.g., copular predicates:



	PreCo F1	OntoNotes F1	$\Delta$
Overall	85.3	64.4	↓20.9 (-24%)
Generics	75.4	0.4	↓75.0 (-99%)



# Evidence

1. Models perform poorly on certain types of coreference
  - a. We evaluate **five types** that generally differ between annotations, across **four datasets**:
    - i. Nested Mentions
    - ii. Generic Mentions
    - iii. Compound Modifiers
    - iv. Copular Predicates
    - v. Pronominal Anaphors



# Evidence


1. Models perform poorly on certain types of coreference
2. Errors correlate more with training set than model architecture



# Evidence

1. Models perform poorly on certain types of coreference
2. Errors correlate more with training set than model architecture

E.g., correlation of errors for models evaluated on PreCo:



$T5_{ON}$	1.00				
WLC	0.89	1.00			
LingMess	0.90	0.90	1.00		
$T5_{PC}$	0.33	0.31	0.32	1.00	
$T5_{WG}$	0.08	0.09	0.08	0.05	1.00
	$T5_{ON}$	WLC	LingMess	$T5_{PC}$	$T5_{WG}$

# Conclusion

1. Failures to generalize are correlated with differences in operationalizations of coreference
2. Valid measurements of model generalization require resolving inconsistencies between operationalizations

