Towards a better understanding of neural relation extraction

Christoph Alt

October 2020

Relation extraction

The measures include Aerolineas's domestic subsidiary, Austral.

Relation extraction

The measures include $\underbrace{\text{Aerolineas's}}_{\text{organization}}$ domestic subsidiary, $\underbrace{\text{Austral}}_{\text{organization}}$.

Relation extraction

The measures include $\underbrace{\text{Aerolineas}}_{\text{organization}}$'s domestic subsidiary, $\underbrace{\text{Austral}}_{\text{organization}}$.

Relation extraction

?

The measures include $\underbrace{\text{Aerolineas}}_{\text{organization}}$'s domestic subsidiary, $\underbrace{\text{Austral}}_{\text{organization}}$.

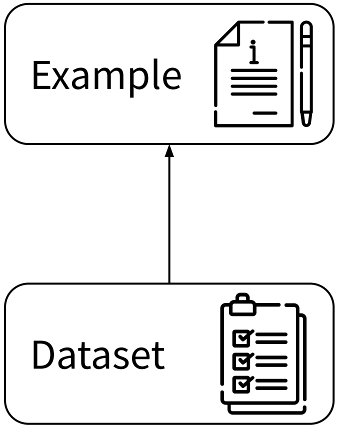
Relation extraction

subsidiary

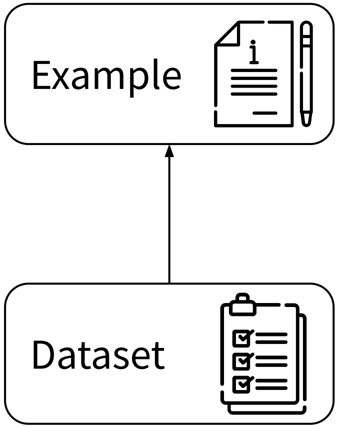
The measures include Aerolineas's domestic subsidiary, Austral.

organization organization

Neural relation extraction

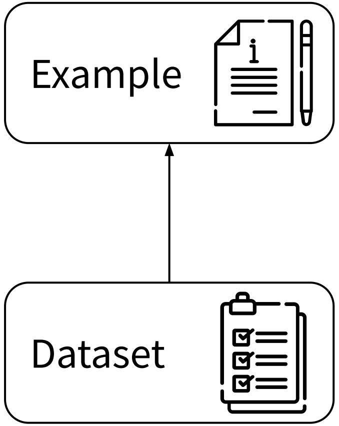


Neural relation extraction



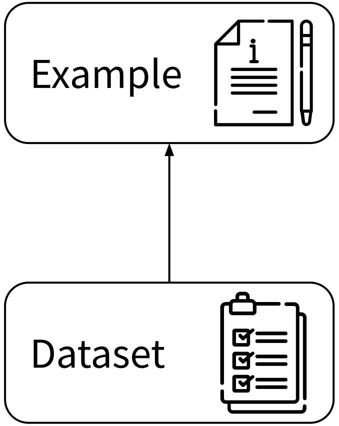
The measures include Aerolineas’s domestic subsidiary, Austral.

Neural relation extraction



The measures include $\underbrace{\text{Aerolineas}}_{\text{head}}$’s domestic subsidiary, $\underbrace{\text{Austral}}_{\text{tail}}$.

Neural relation extraction



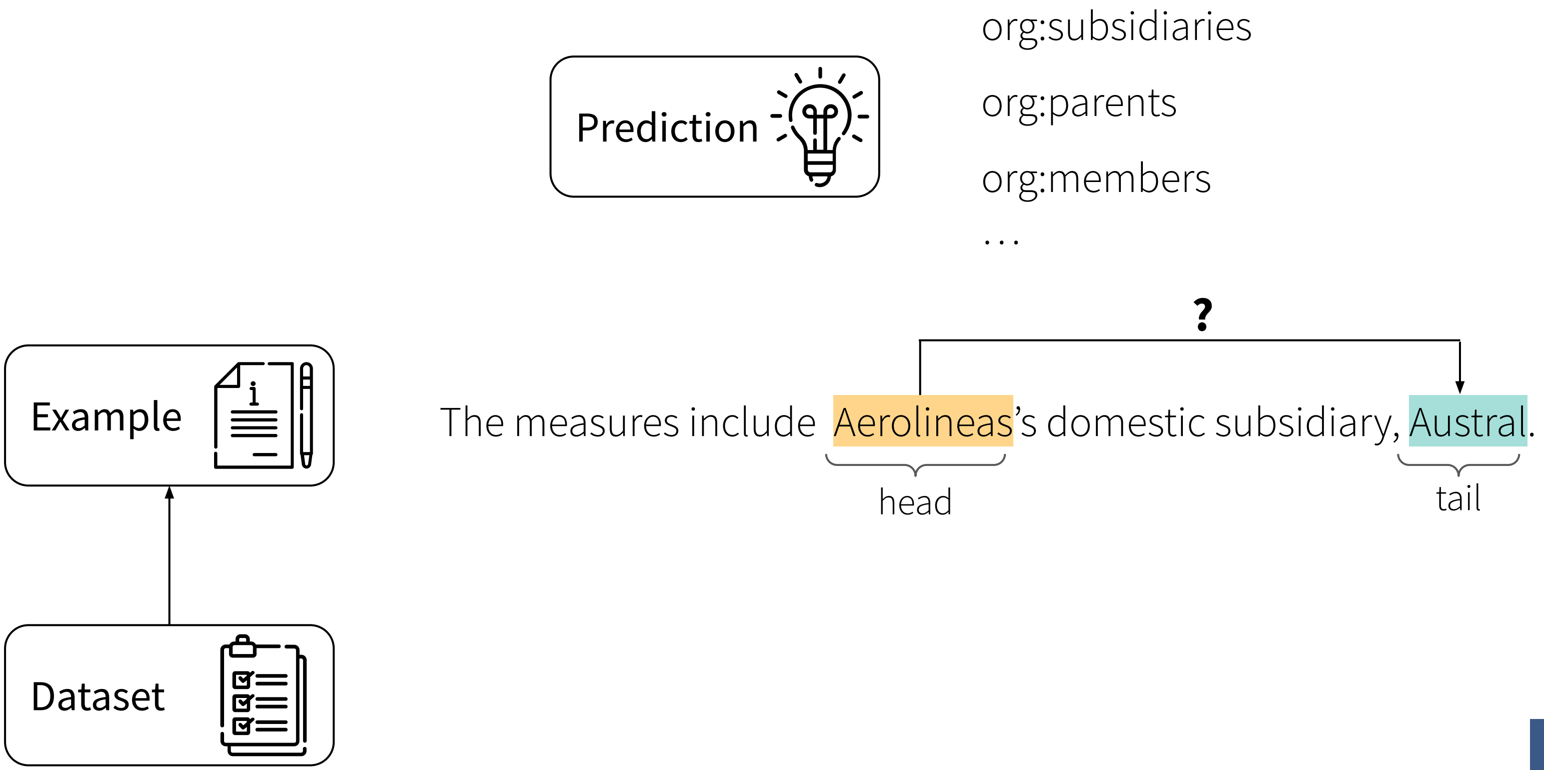
?

The measures include Aerolineas's domestic subsidiary, Austral.

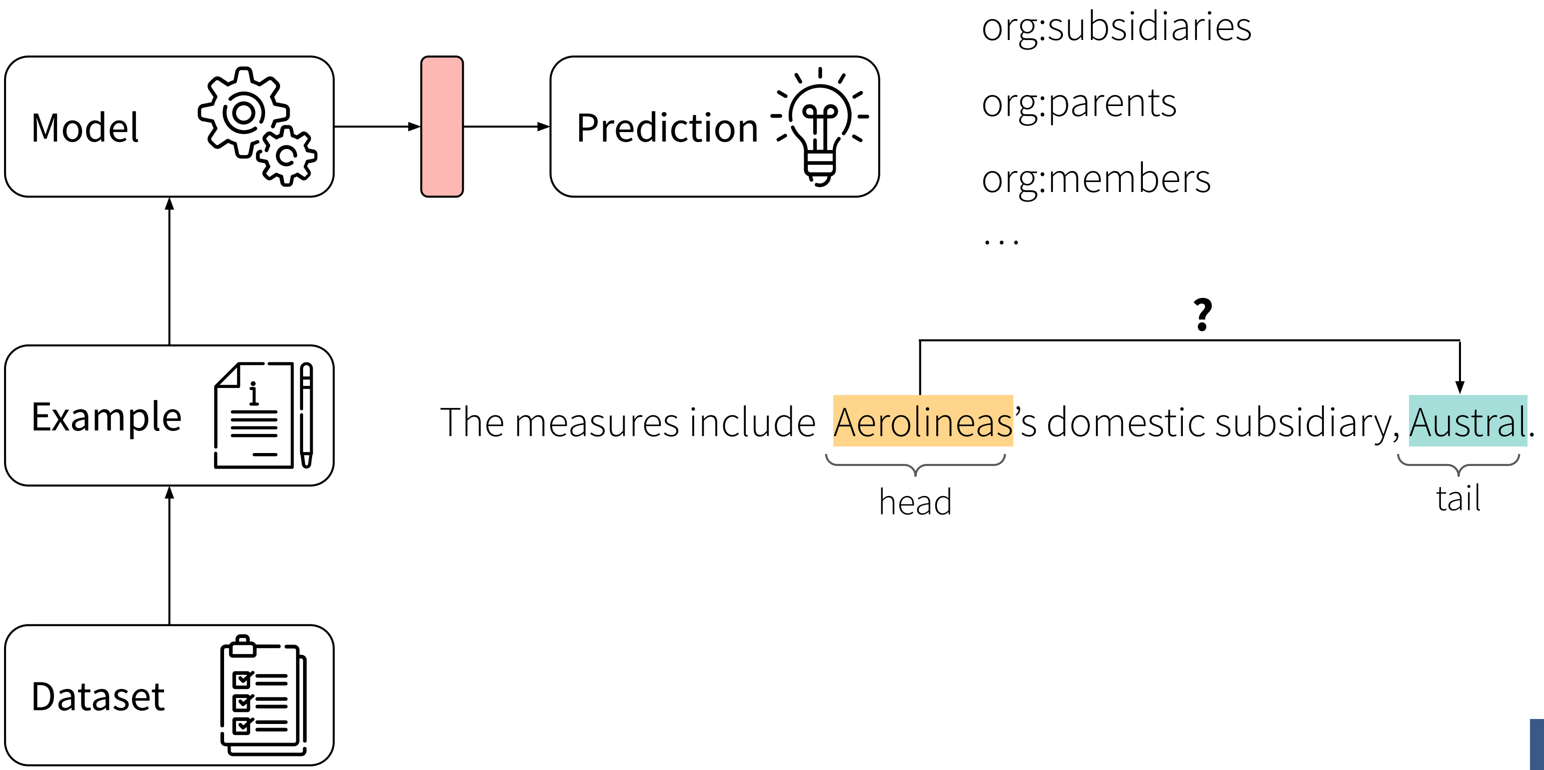
head

tail

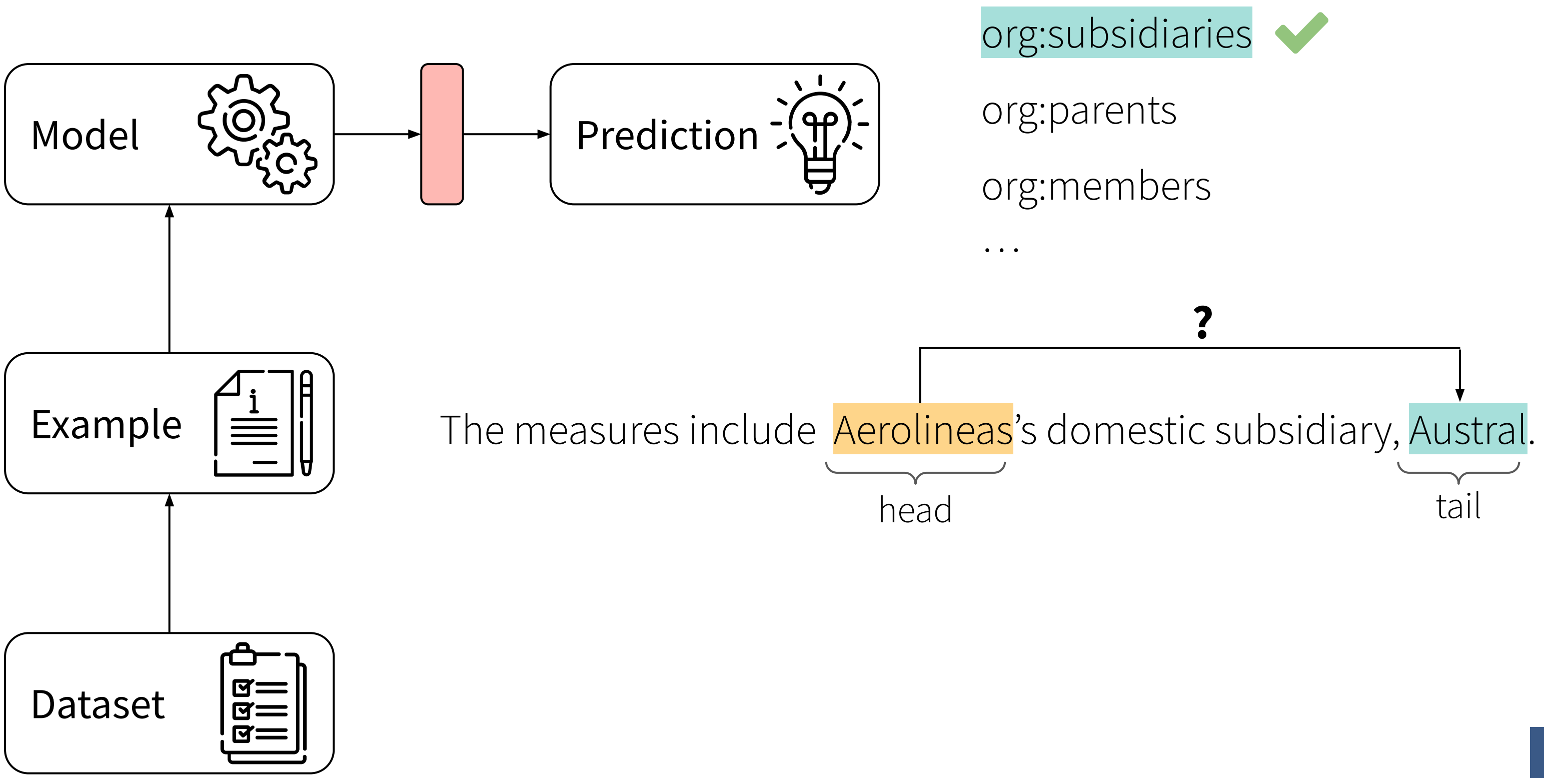
Neural relation extraction



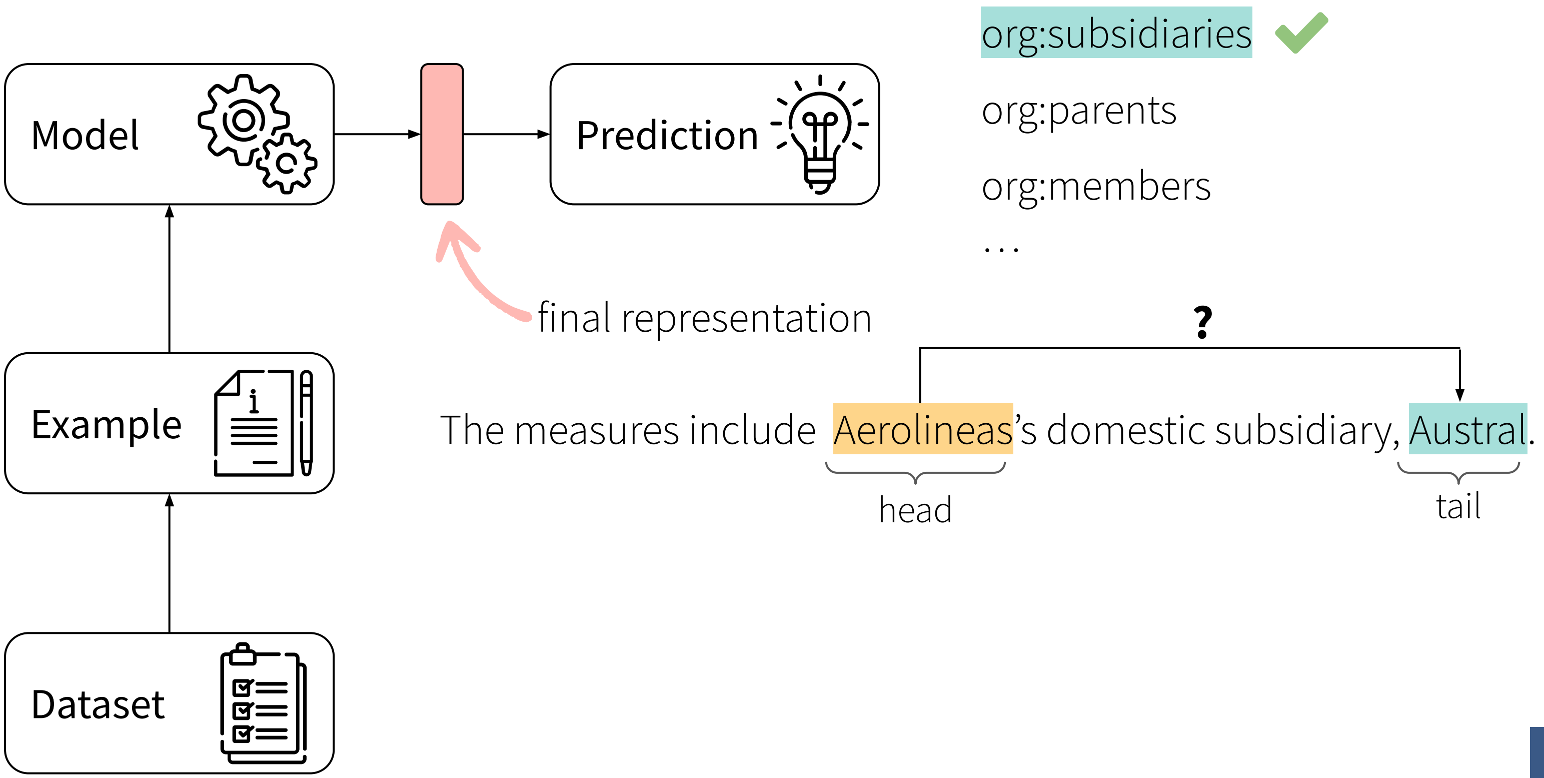
Neural relation extraction



Neural relation extraction

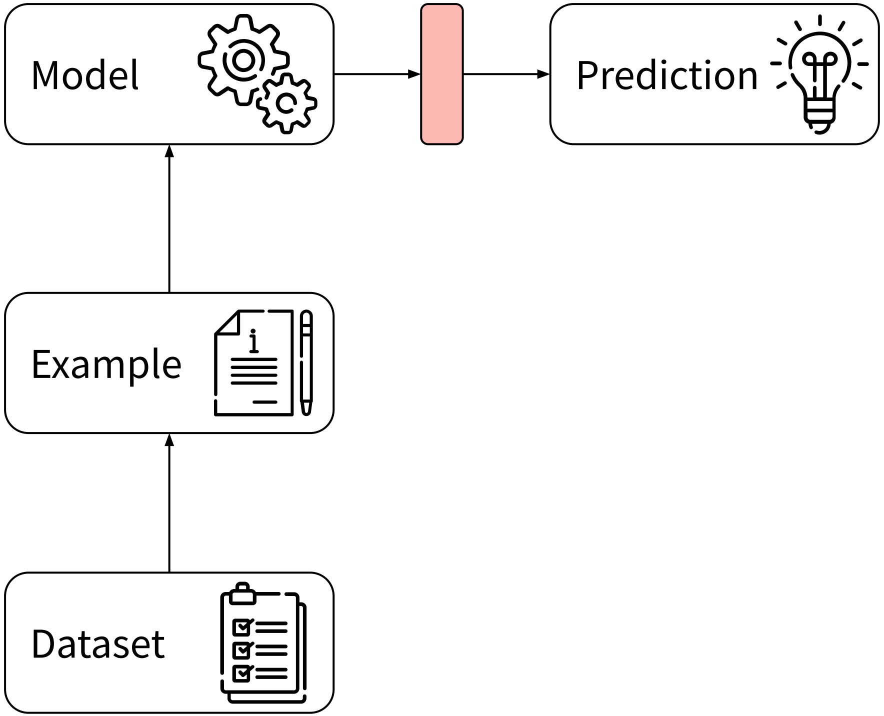


Neural relation extraction

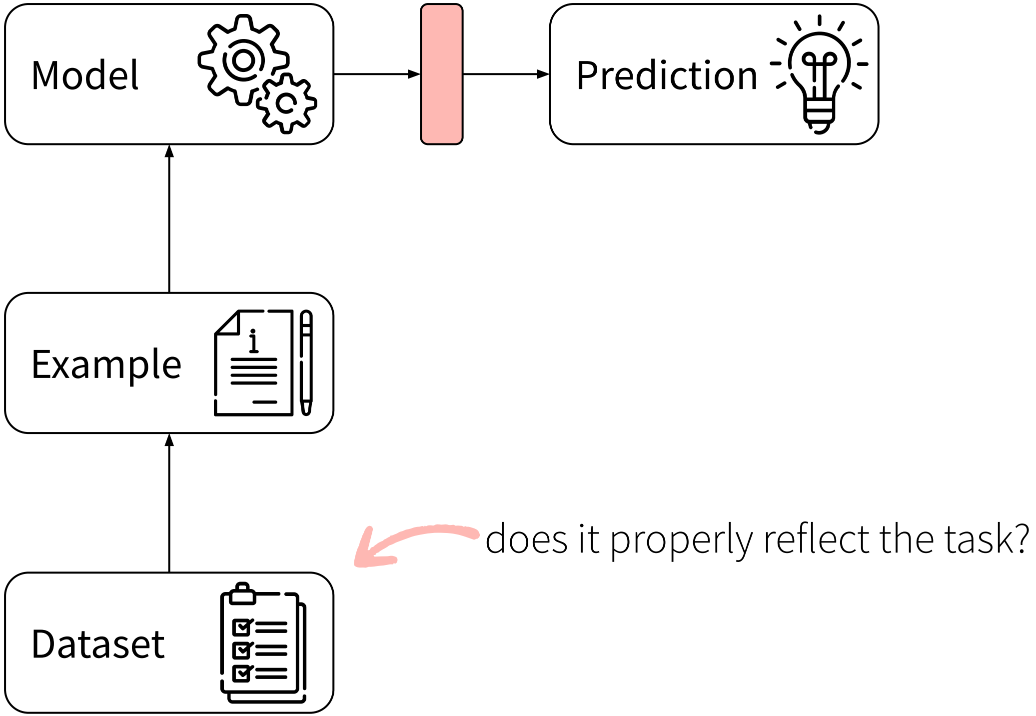


How do we get a better understanding of neural relation extraction?

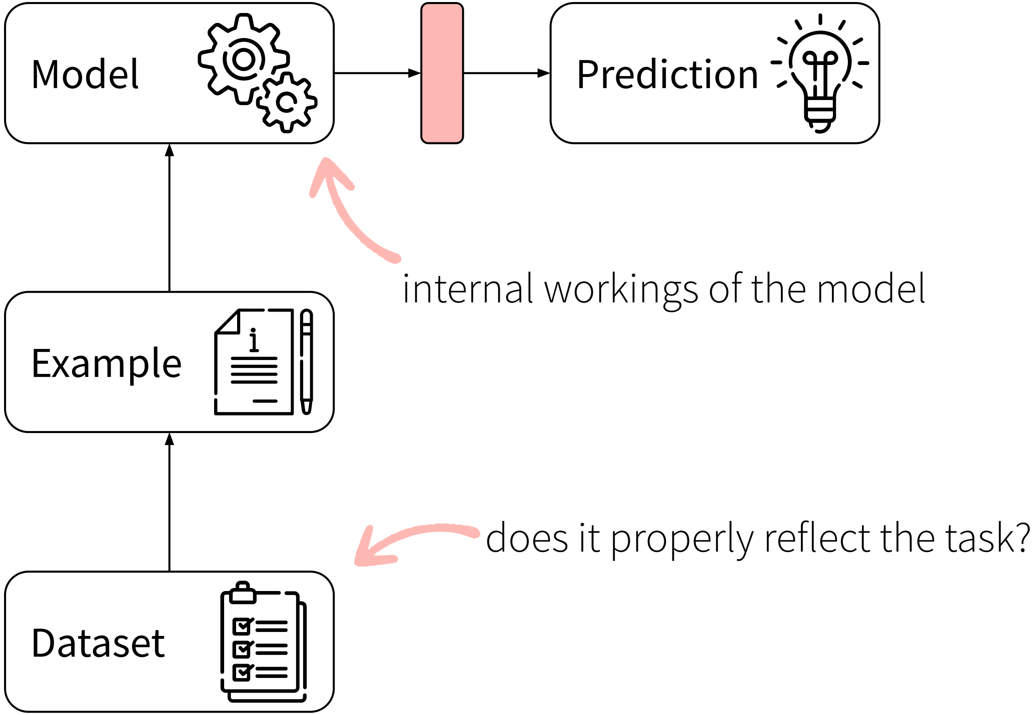
Understanding neural relation extraction



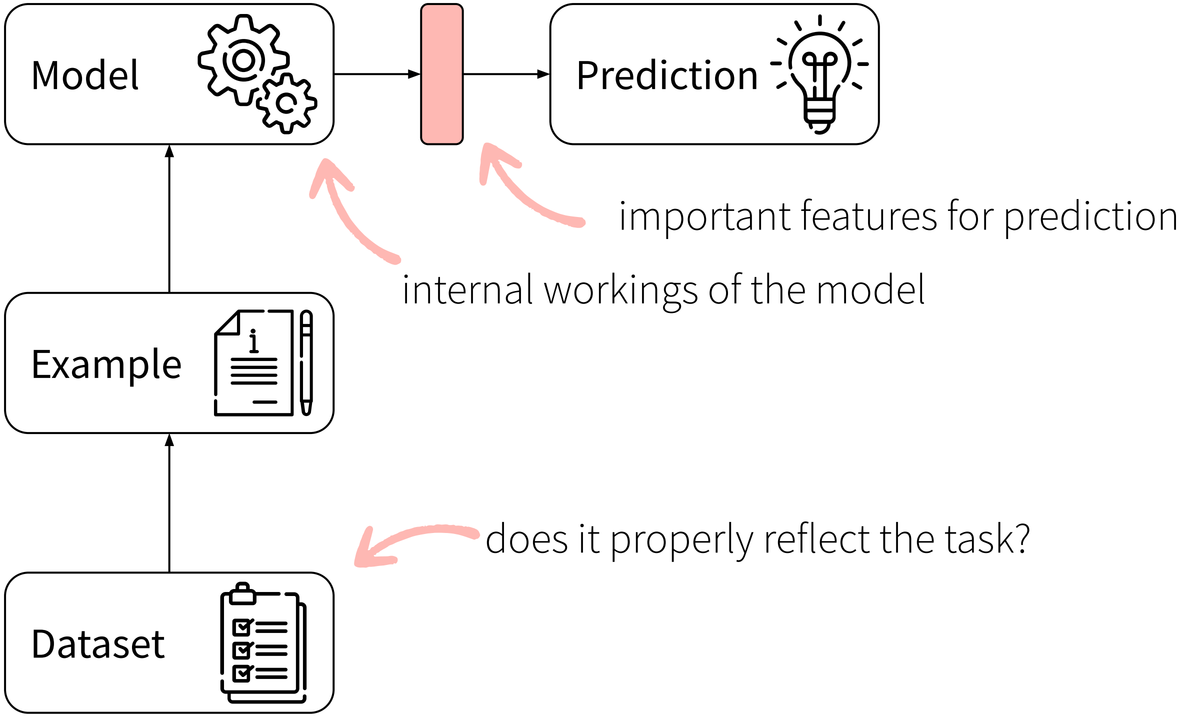
Understanding neural relation extraction



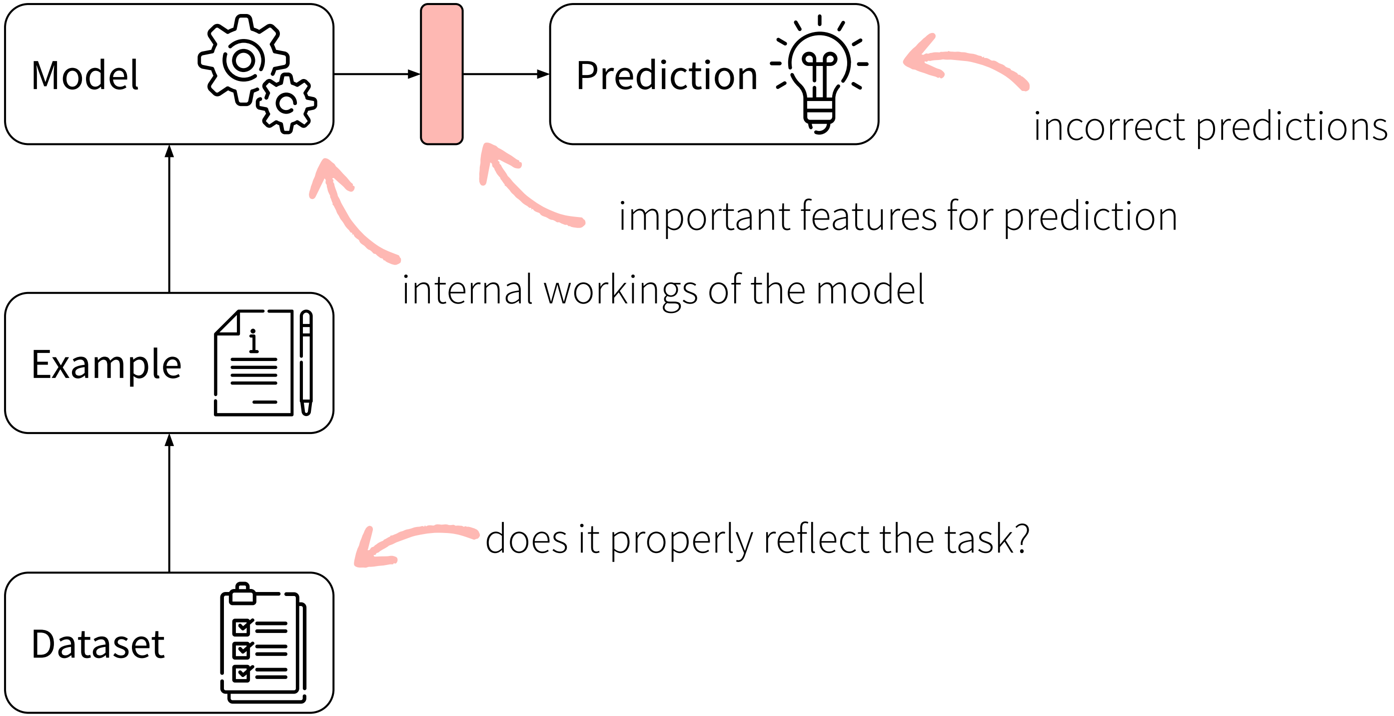
Understanding neural relation extraction



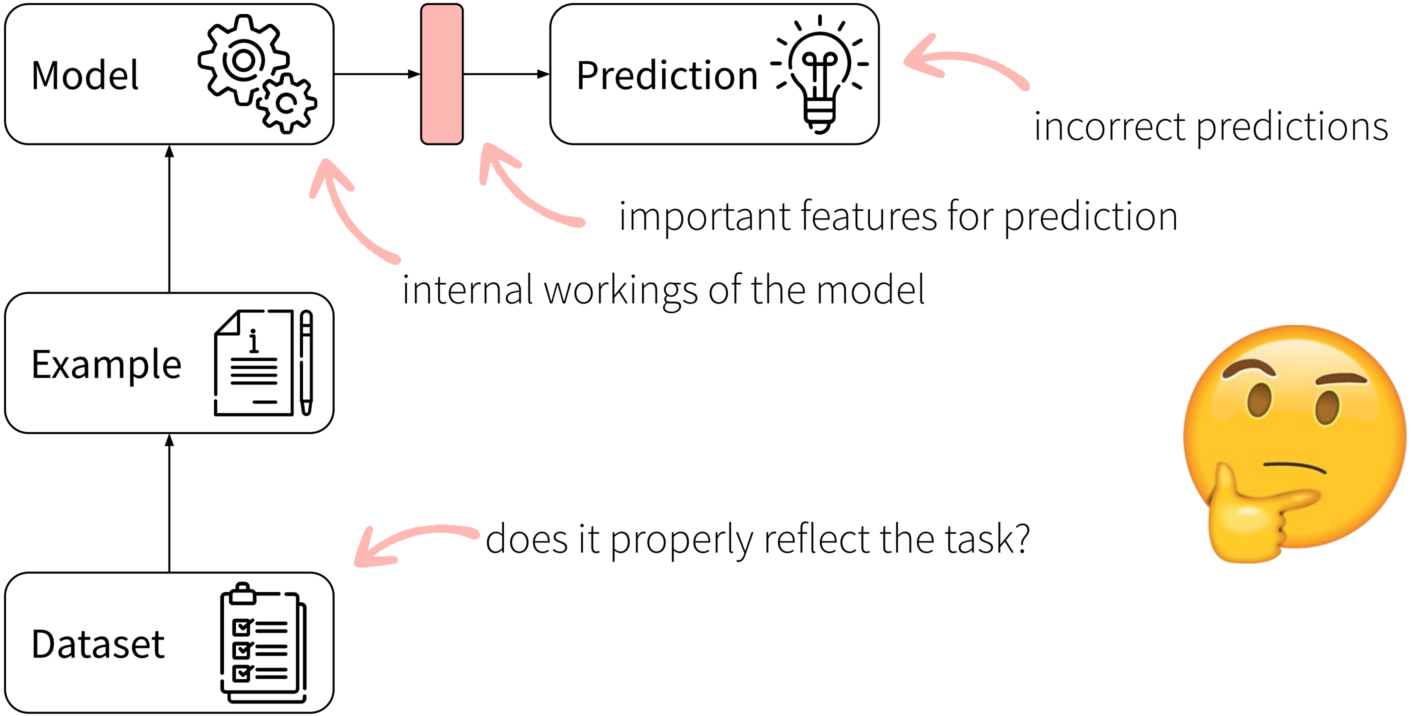
Understanding neural relation extraction



Understanding neural relation extraction



Understanding neural relation extraction



In this talk

1. What linguistic aspects of the input do neural relation extraction models focus on?

In this talk

1. What linguistic aspects of the input do neural relation extraction models focus on?
2. Where do neural relation extraction models fail, and why?

1. **What linguistic aspects of the input do neural relation extraction models focus on?**
2. Where do neural relation extraction models fail, and why?

Probing Linguistic Features of Sentence-Level Representations in Neural Relation Extraction. Christoph Alt, Aleksandra Gabryszak and Leonhard Hennig. ACL 2020

What properties are important to relation extraction?

The measures include $\underbrace{\text{Aerolineas}}_{\text{head}}$'s domestic subsidiary, $\underbrace{\text{Austral}}_{\text{tail}}$.

What properties are important to relation extraction?

The measures include $\underbrace{\text{Aerolineas}}_{\text{head}}$'s domestic subsidiary, $\underbrace{\text{Austral}}_{\text{tail}}$.

entity type?

What properties are important to relation extraction?

The measures include $\underbrace{\text{Aerolineas}}_{\text{head}}$'s domestic subsidiary, $\underbrace{\text{Austral}}_{\text{tail}}$.

argument order?

entity type?

What properties are important to relation extraction?

The measures include $\underbrace{\text{Aerolineas}}_{\text{head}}$'s domestic subsidiary, $\underbrace{\text{Austral}}_{\text{tail}}$.

argument order?

distance between arguments?

entity type?

What properties are important to relation extraction?

sentence structure?

The measures include Aerolineas's domestic subsidiary, Austral.

head

tail

argument order?

distance between arguments?

entity type?

What properties are important to relation extraction?

sentence structure?

The measures include $\underbrace{\text{Aerolineas}}_{\text{head}}$'s domestic subsidiary, $\underbrace{\text{Austral}}_{\text{tail}}$.

argument order?

distance between arguments?

entity type?

Do representations contain any of these properties?

sentence structure?

The measures include Aerolineas's domestic subsidiary, Austral.

head

tail

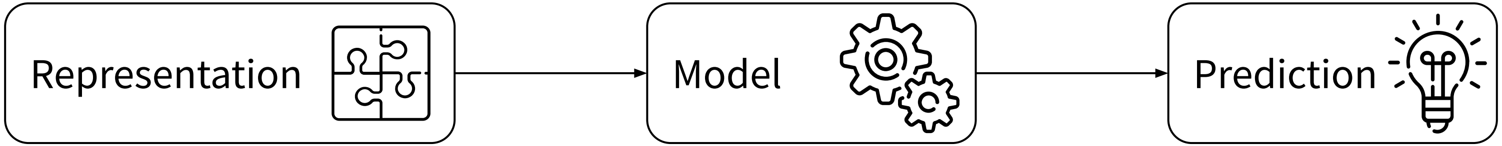
argument order?

distance between arguments?

entity type?

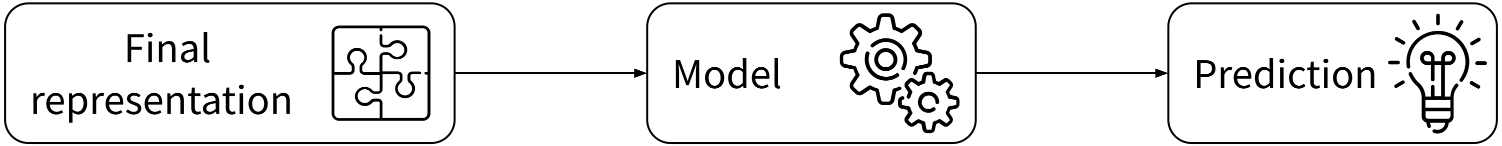
Do representations contain any of these properties?

Probing tasks



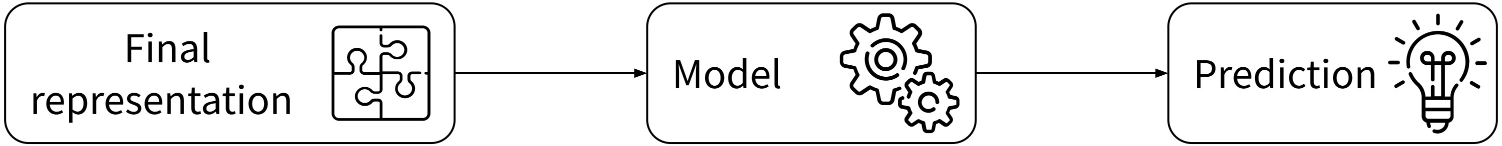
- Probing task, diagnostic classifier or auxiliary prediction task [Adi et al., 2017, Conneau et al., 2018]
  - Simple classification task, classifier trained on representations
  - Performance measures how well the information is encoded
    - → Assumption: Information is used for model prediction

Probing tasks



Probing tasks

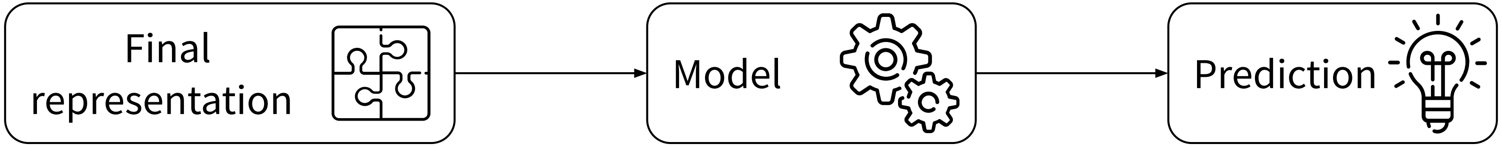
Model architectures



- Bag of embeddings
- CNN
- GCN (graph conv.)
- (Bi-) LSTM
- Self-attention

Probing tasks

Supporting linguistic features



Bag of embeddings

CNN

GCN (graph conv.)

(Bi-) LSTM

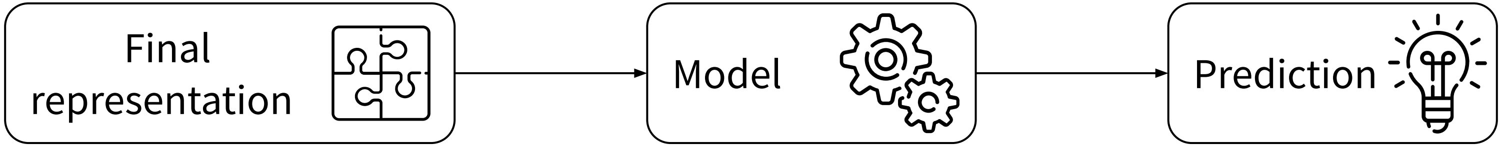
Self-attention

Entity masking

Contextual word represent.

Probing tasks

Tasks



- Bag of embeddings
- CNN
- GCN (graph conv.)
- (Bi-) LSTM
- Self-attention

Entity masking

Contextual word represent.

- Surface properties
- Syntactic properties
- Semantic properties

Tasks

Surface properties

- Sentence length
- Argument distance
- Named entity between arguments

Tasks

Surface properties

- Sentence length
- Argument distance
- Named entity between arguments

Syntactic properties

- Dependency tree depth
- Shortest dependency path tree depth
- Argument order
- POS of tokens to the left and right of {head, tail}

Tasks

Surface properties

- Sentence length
- Argument distance
- Named entity between arguments

Syntactic properties

- Dependency tree depth
- Shortest dependency path tree depth
- Argument order
- POS of tokens to the left and right of {head, tail}

Semantic properties

- Named entity type of {head, tail}
- Grammatical role of {head, tail}

Experiment Setup

Probing task dataset:

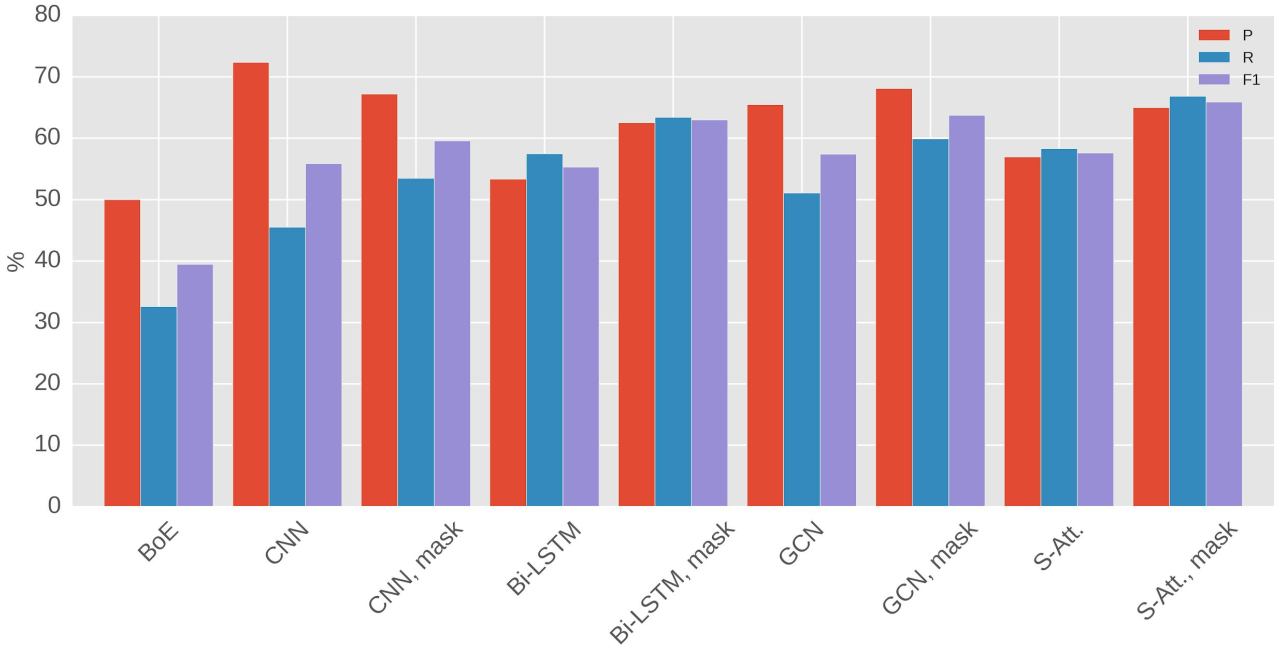
- Collect sentences from TACRED [Zhang et al., 2017] and SemEval 2010 Task 8 [Hendrickx et al., 2010]
- Assign probing task label
  - syntactic and semantic probing tasks labels via Stanford CoreNLP [Manning et al., 2014]

Evaluation approach:

- Train relation extraction model, e.g., on TACRED
- Evaluate accuracy of probing task model trained on final representation

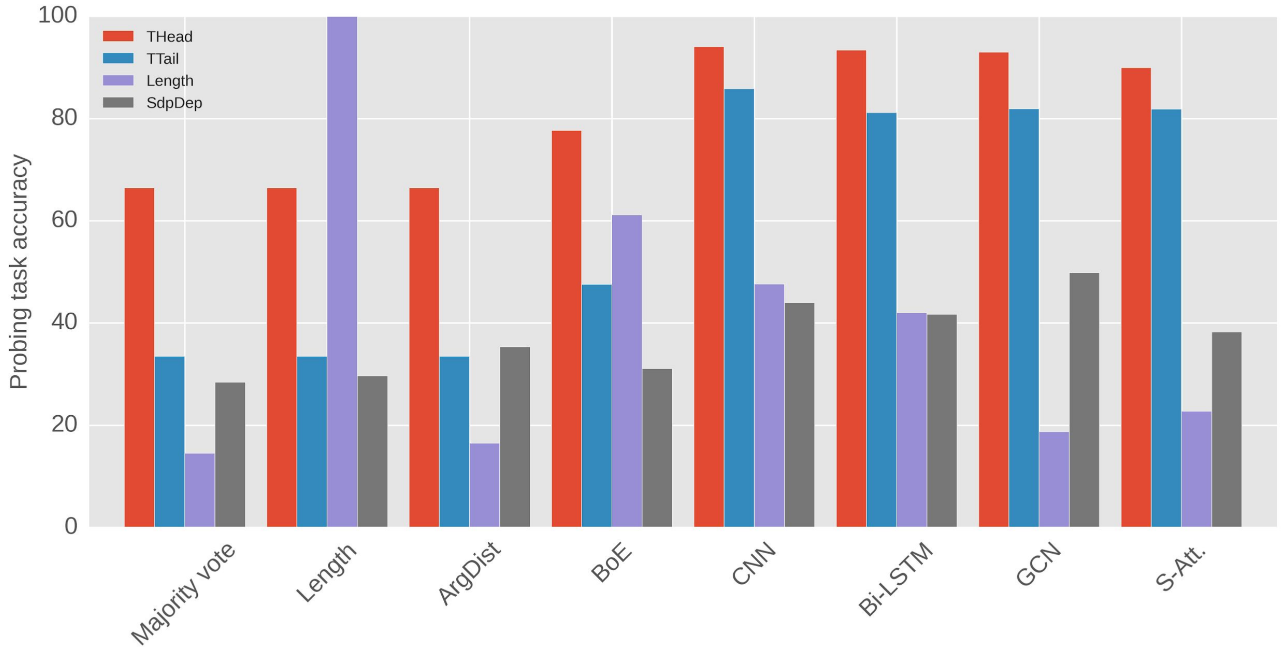
Results

Overall relation extraction performance



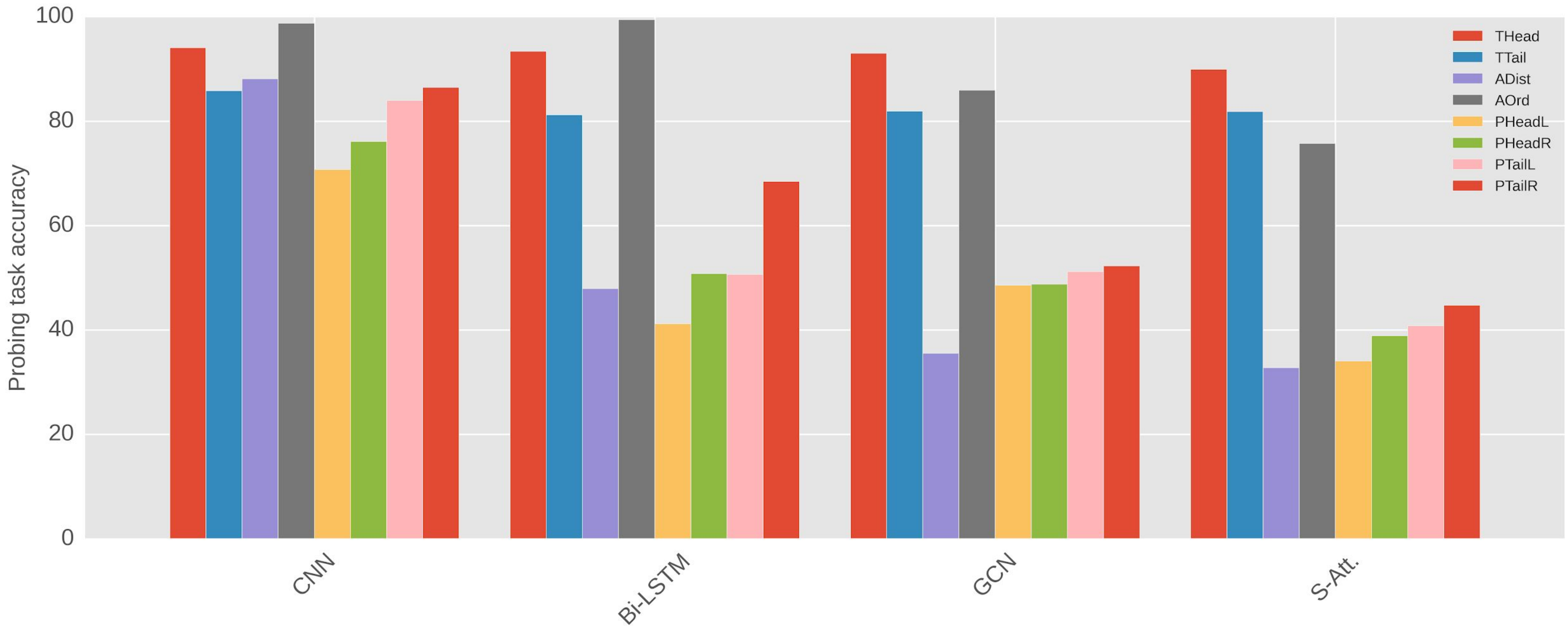
Results

General probing task performance



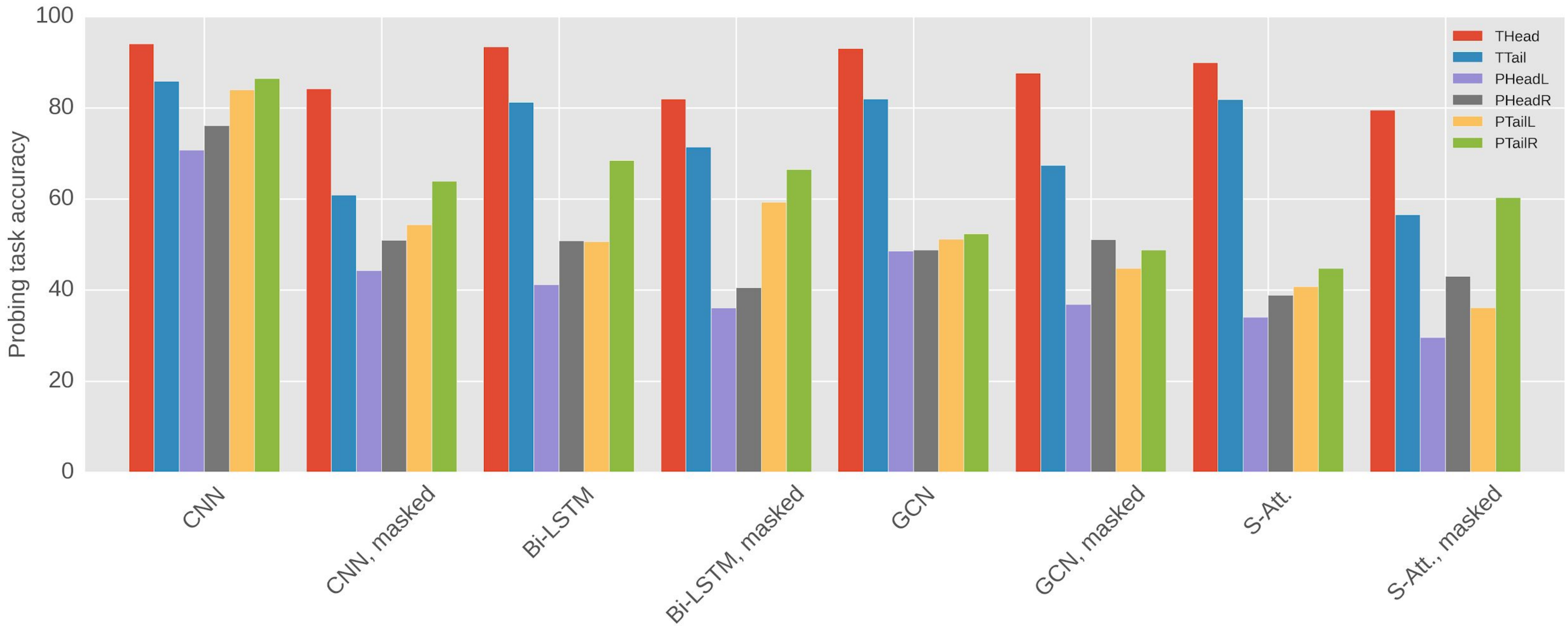
Results

Neural network architecture



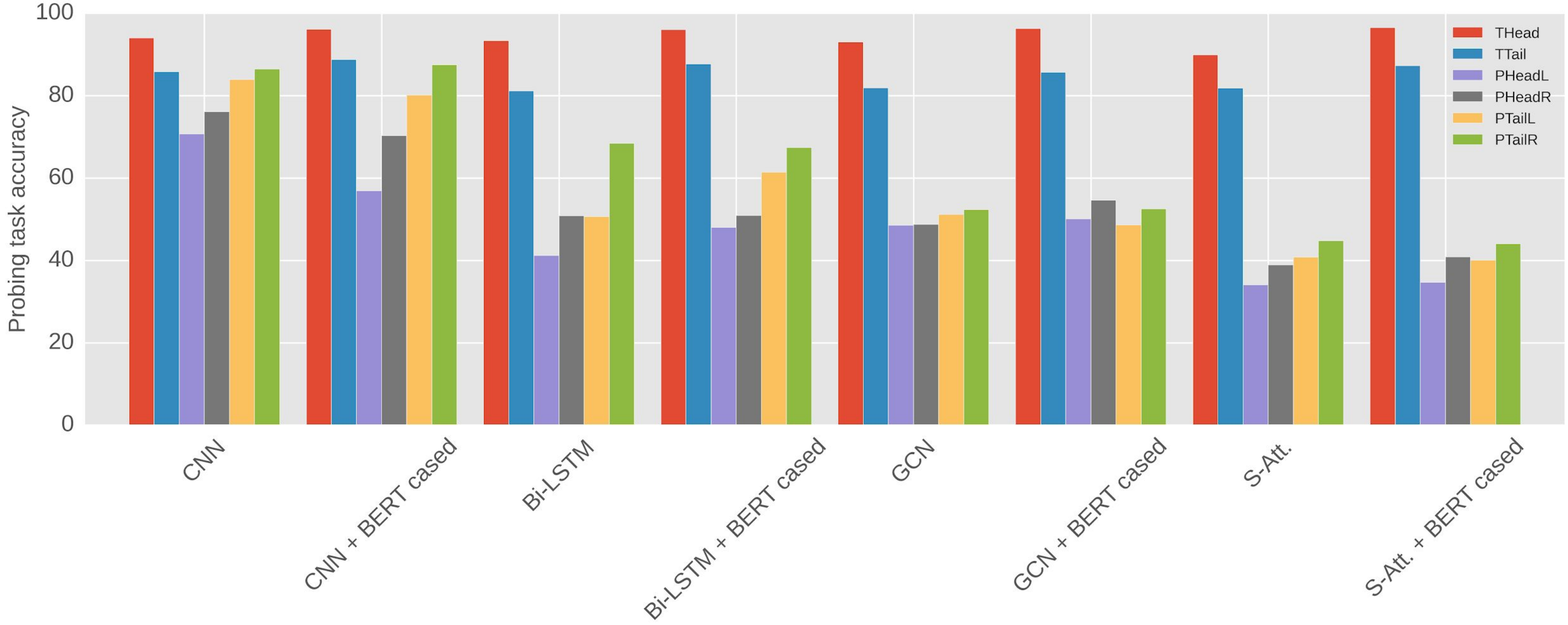
Results

Entity masking



Results

Contextual word representations



Summary

- Extensive evaluation showed that
    - self-attentive encoders are well suited for RE
    - but perform lower on probing tasks
    - bias induced by different architectures is reflected in probing task performance
        - e.g., distance and dependency related tasks
- However, probing task performance *not correlated with RE performance*

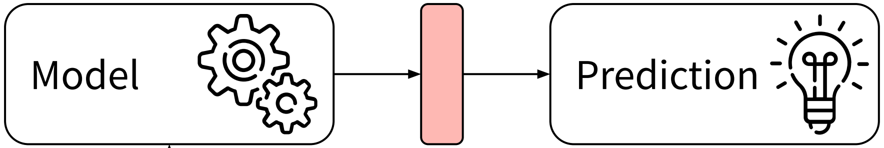
Software libraries:

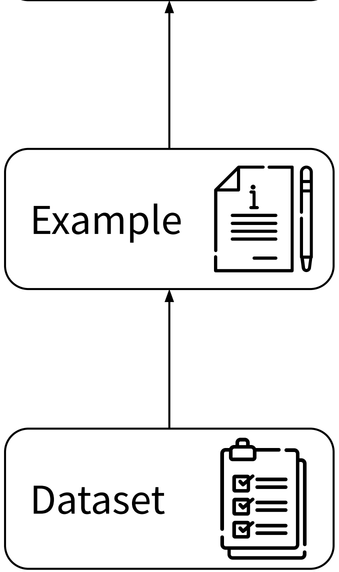
- REval: framework to develop and evaluate probing tasks for neural RE, based on SentEval [Conneau and Kiela, 2018]
- RelEx: binary RE framework based on AllenNLP [Gardner et al., 2017]

1. What linguistic aspects do neural relation extraction models focus on?
2. **Where do neural relation extraction models fail, and why?**

TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task.
Christoph Alt, Aleksandra Gabryszak and Leonhard Hennig. ACL 2020

Model errors

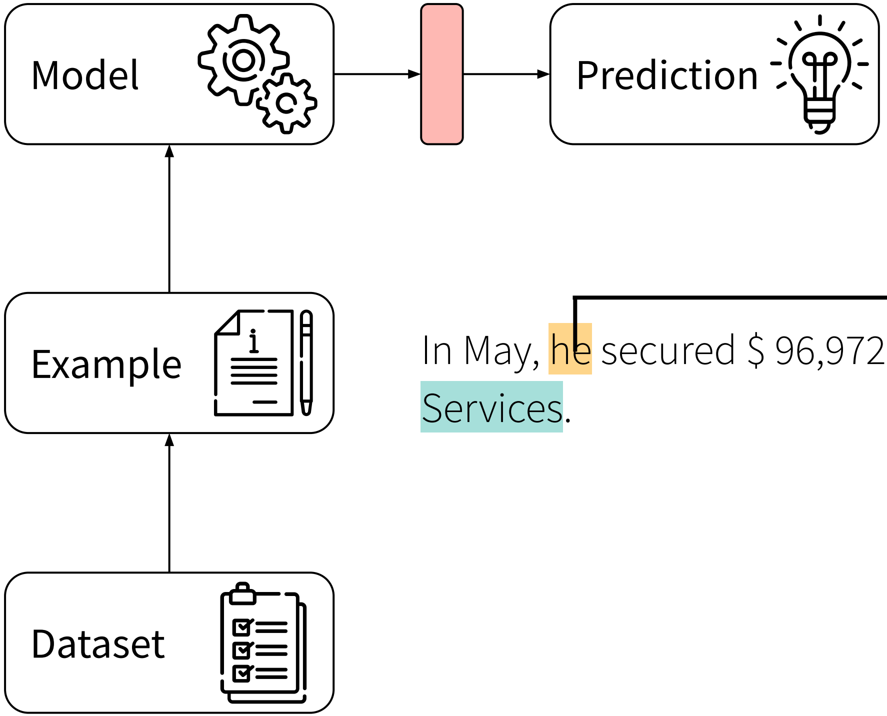




?

In May, he secured $ 96,972 in working capital from GE Healthcare Financial Services.

Model errors



per:employee_of

org:parents

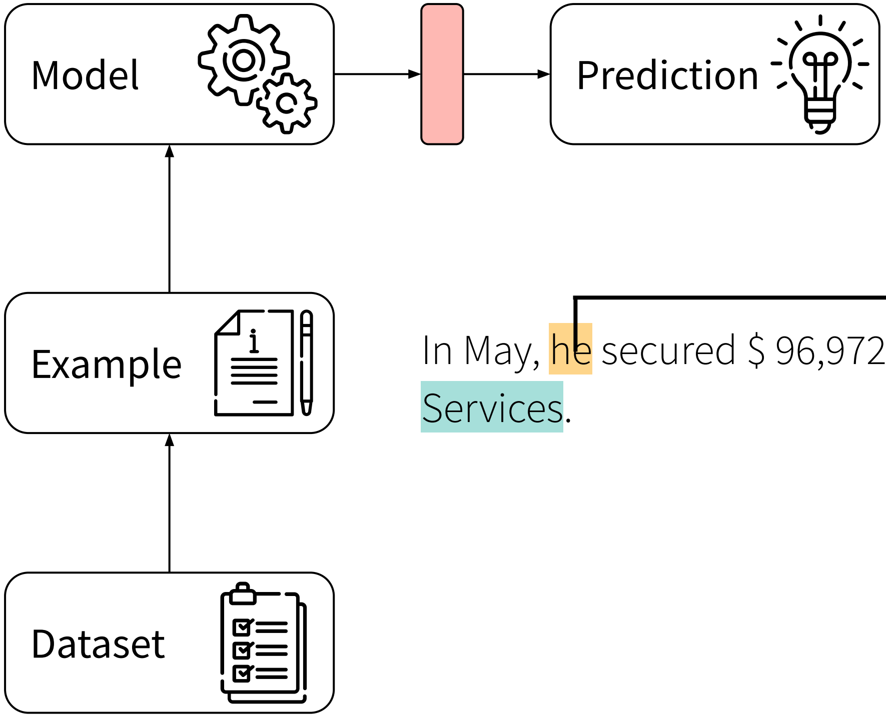
org:members

...

?

In May, he secured $ 96,972 in working capital from GE Healthcare Financial Services.

Model errors



per:employee_of

org:parents

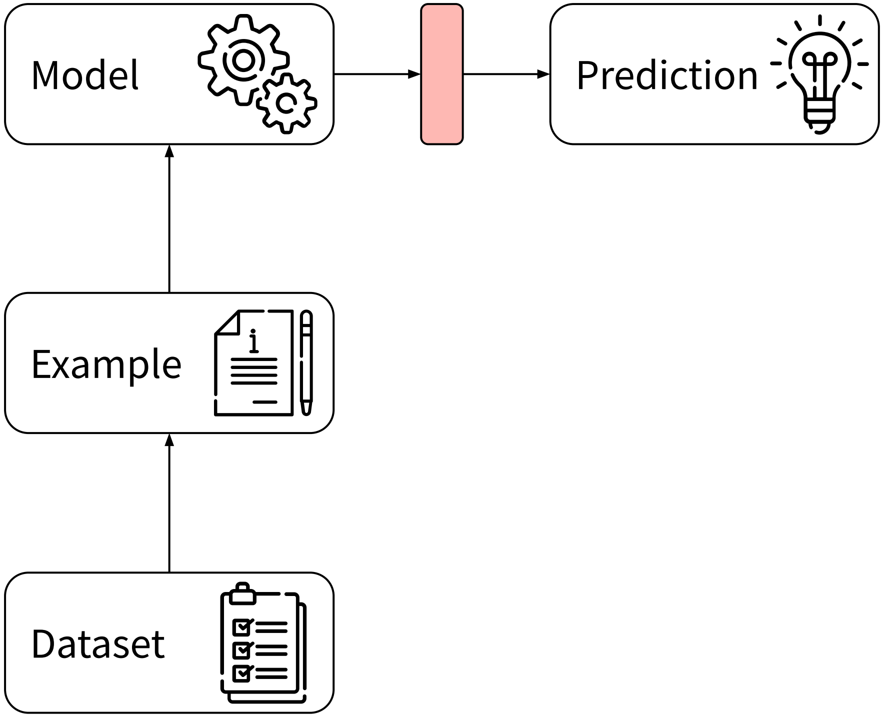
org:members

. . .

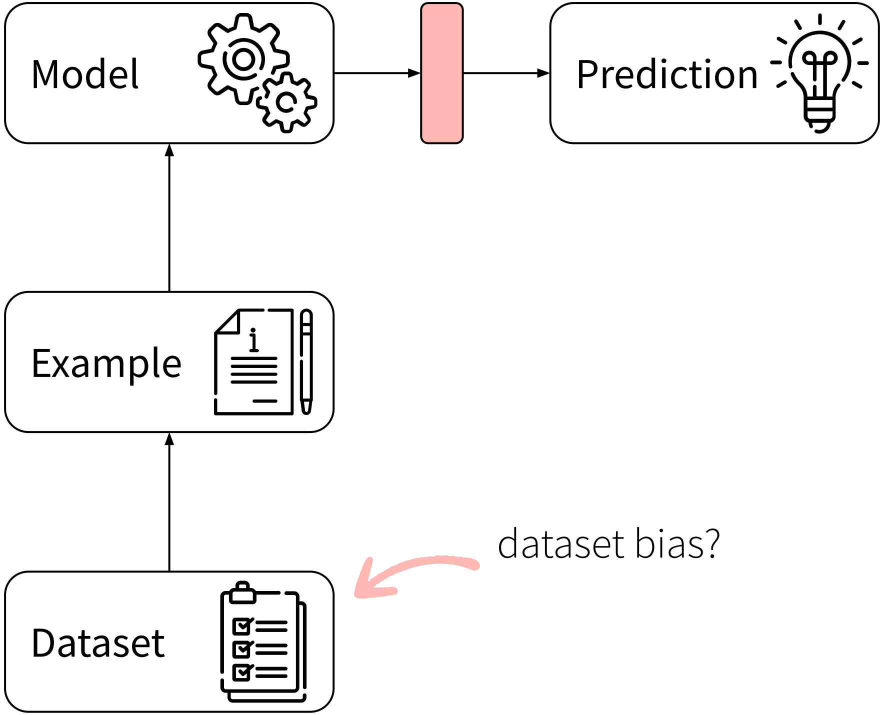
?

In May, he secured $ 96,972 in working capital from GE Healthcare Financial Services.

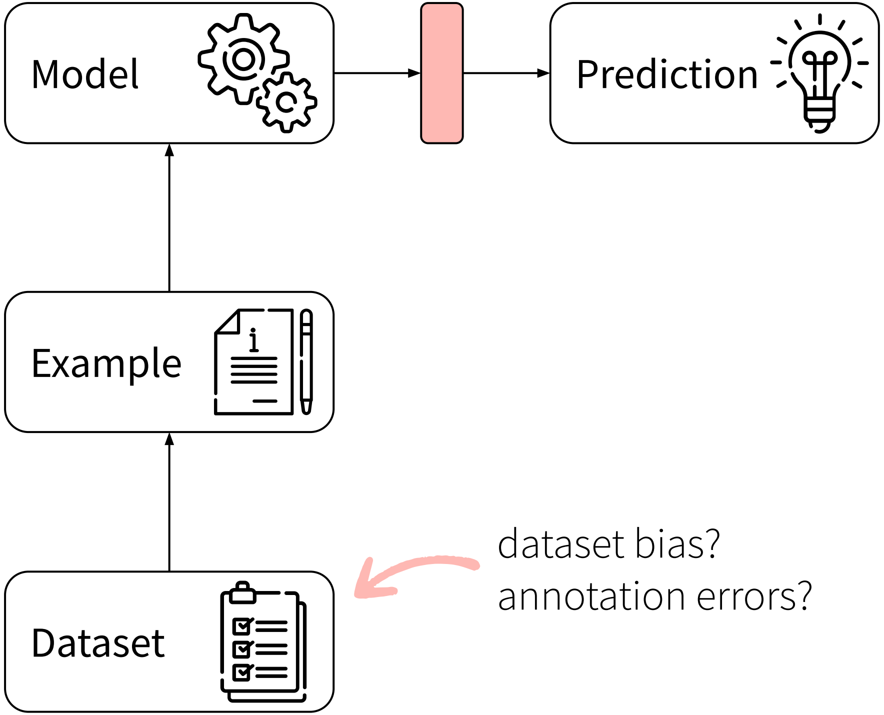
What is causing model errors?



What is causing model errors?

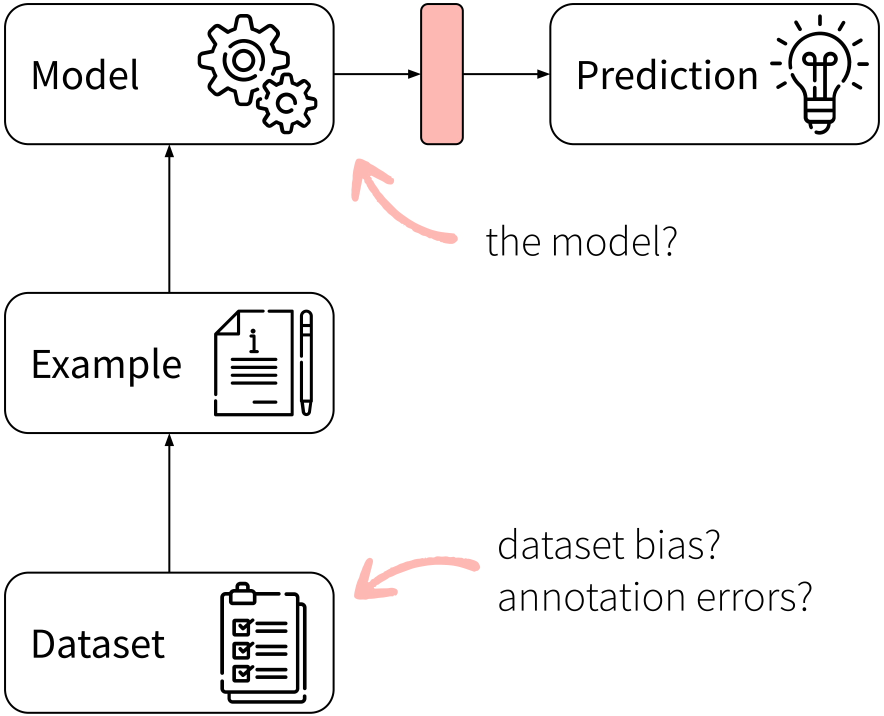


What is causing model errors?

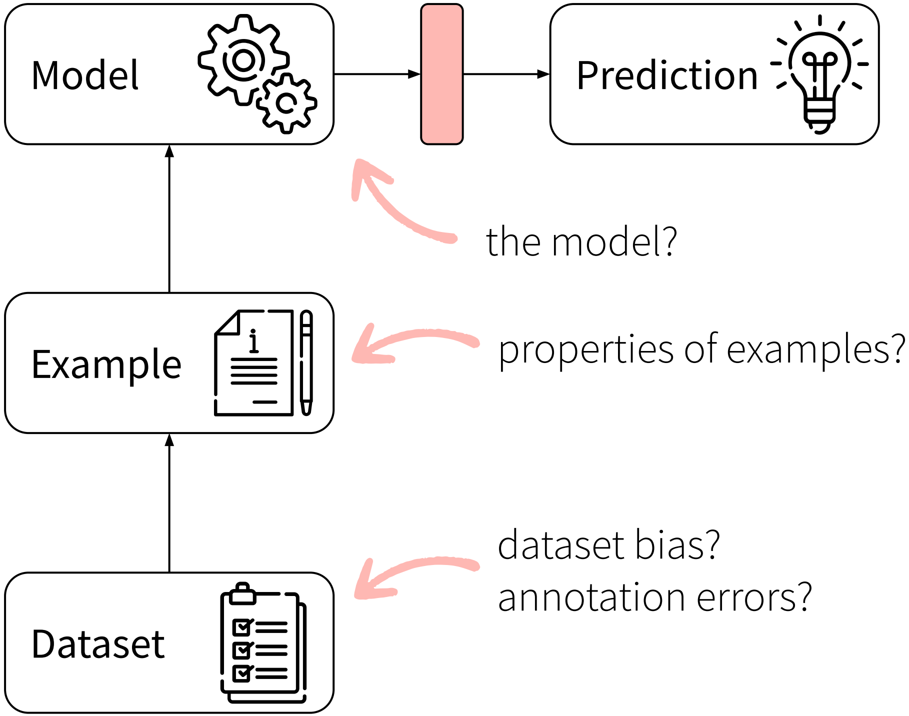


dataset bias?
annotation errors?

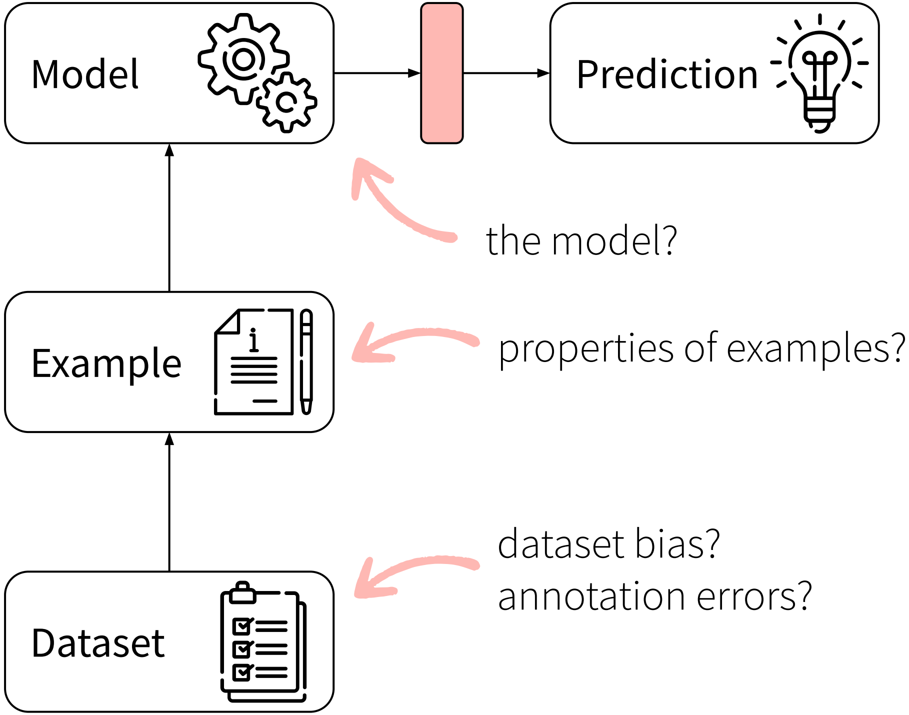
What is causing model errors?



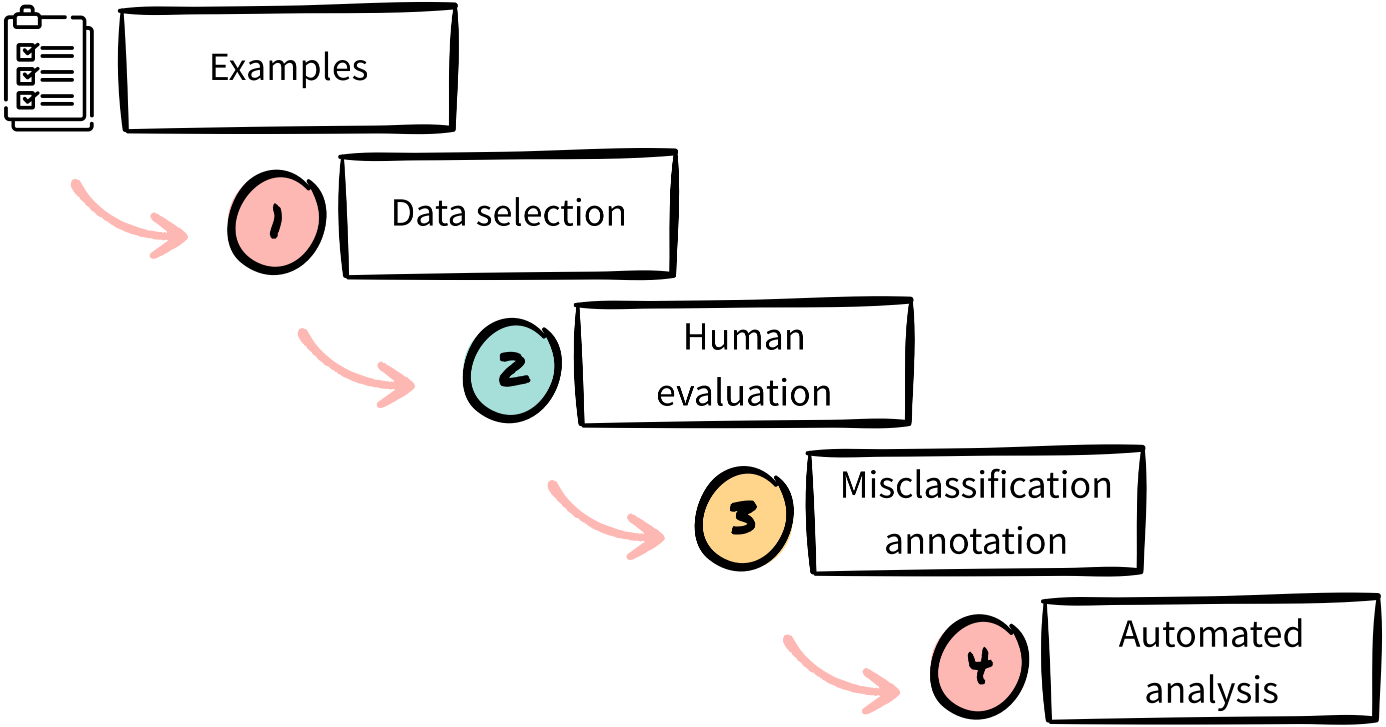
What is causing model errors?



What is causing model errors?



General approach



Evaluation

Examples

TACRED
(106k examples, 41 relations)

1 Data selection

2 Human evaluation

3 Misclassification annotation

4 Automated analysis

Evaluation

Data selection and human evaluation

Approach:

- Rank each example according to evidence from 49 different RE model predictions
- Select examples for manual evaluation
  - *Challenging* → misclassified by at least half of the models
  - *Control* → Correctly classified by at least 39 models
- Manual re-annotation of selected examples

Evaluation

Examples

TACRED
(106k examples, 41 relations)

1 Data selection

2 Human evaluation

3 Misclassification annotation

4 Automated analysis

Evaluation

Examples

TACRED
(106k examples, 41 relations)

1. Data selection

based on errors of 49 different RE models

2. Human evaluation

3. Misclassification annotation

4. Automated analysis

Evaluation

Examples — TACRED (106k examples, 41 relations)

1. Data selection — based on errors of 49 different RE models
2. Human evaluation — re-annotation of 5k examples (dev and test)
3. Misclassification annotation
4. Automated analysis

Results (1): label error analysis

	Dev		Test	
	Challenging	*Control*	*Challenging*	*Control*
# Examples (# positive)	3,088 (1,987)	567 (547)	1,923 (1,333)	427 (407)
# Revised (# positive)	1,610 (976)	46 (46)	960 (630)	38 (38)
# Revised (% positive)	**52.1 (49.1)**	**8.1 (8.4)**	**49.9 (47.3)**	**8.9 (9.3)**

- Approx. 5k challenging examples re-annotated
- Approx. 50% of challenging examples were revised (relabeled)
- Only 8% of examples in control were revised

Results (1): label error analysis

	Dev		Test	
IAA	H1,H2	H,C	H1,H2	H,C
Challenging	0.78	0.43	0.85	0.44
Control	0.87	0.95	0.94	0.96
All	0.80	0.53	0.87	0.55

H1, H2 → agreement between human re-annotators

H, C → Average agreement between re-annotators and crowd

- High inter-annotator agreement (IAA)
- Challenging set more difficult for re-annotators (H1, H2), too
- Moderate agreement between re-annotators and crowd (H, C)
- Typical crowd errors
  - incorrect positive (49%) → revised to “no relation”
  - incorrect negative (36%)

Results (1): label error analysis

Model	Original	Revised
CNN, masked	59.5	66.5
TRE	67.4	75.3
SpanBERT	70.8	78.0
KnowBERT	71.5	79.3

- Approx. 8% absolute improvement in F1 score across all models
- Average score across 49 models from 62.1 to 70.1 F1
- State-of-the-art improved from 71.5 to 79.3 F1

Evaluation

Misclassification annotation

Approach:

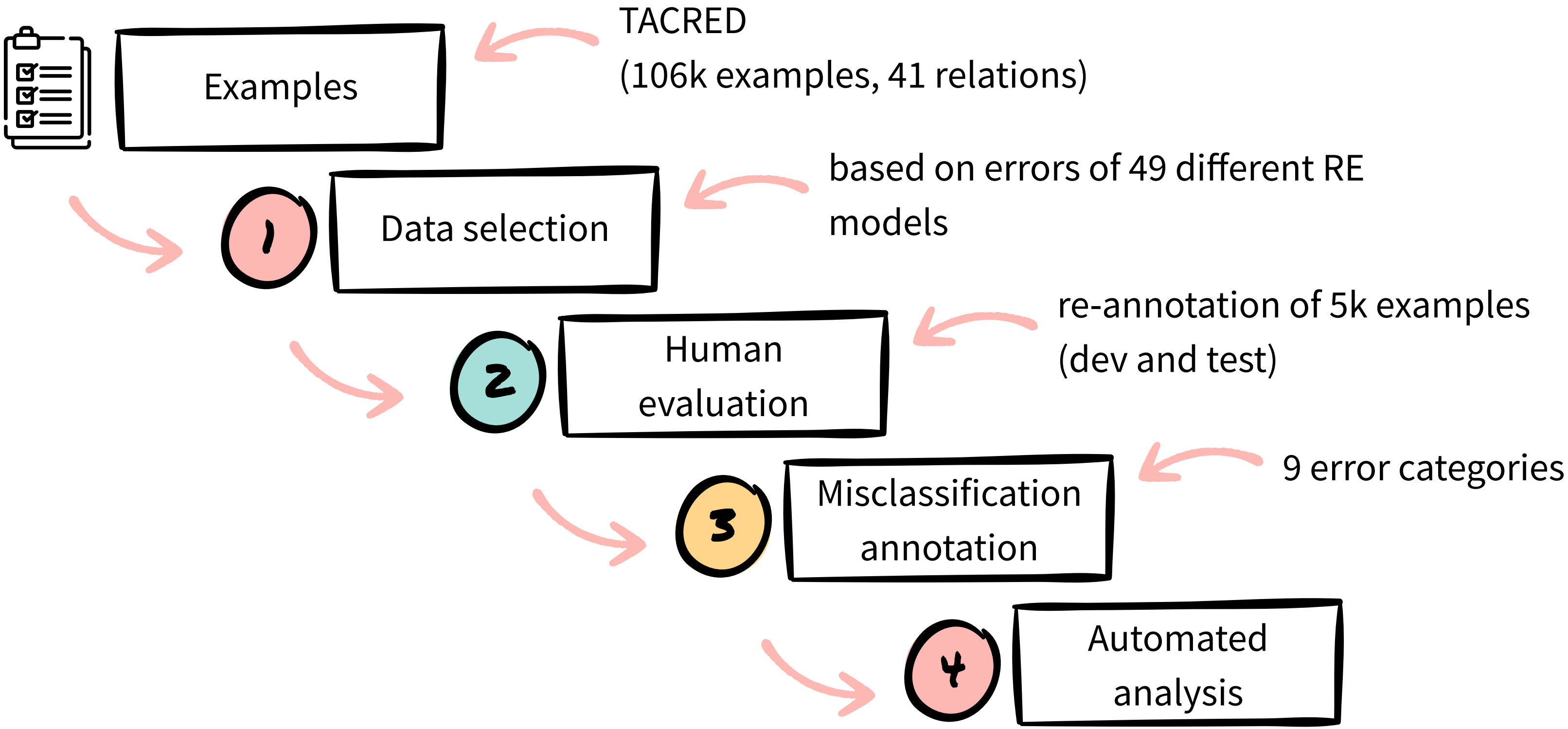
- Explore possible linguistic aspects causing incorrect predictions
    - e.g., entity type errors or distracting phrases
- Iteratively develop error categories
- Annotate each misclassification with category

Evaluation

Examples — TACRED (106k examples, 41 relations)

1. Data selection — based on errors of 49 different RE models
2. Human evaluation — re-annotation of 5k examples (dev and test)
3. Misclassification annotation
4. Automated analysis

Evaluation


Examples
TACRED
(106k examples, 41 relations)
1
Data selection
based on errors of 49 different RE models
2
Human evaluation
re-annotation of 5k examples (dev and test)
3
Misclassification annotation
9 error categories
4
Automated analysis

Results (2): model error categories

Context			
Inverted Args	[Ruben van Assouw]$_{head:per}$, who had been on safari with his 40-year-old father [Patrick]$_{tail:per}$, mother Trudy , 41 , and brother Enzo , 11 .	*per:children*	25
Wrong Args	Authorities said they ordered the detention of Bruno 's wife , [Dayana Rodrigues]$_{tail:per}$, who was found with [Samudio]$_{head:per}$'s baby .	*per:spouse*	109
Ling. Distractor	In May , [he]$_{head:per}$ secured $ 96,972 in working capital from [GE Healthcare Financial Services]$_{tail:org}$.	*per:employ._of*	35
Factuality	[Ramon]$_{head:per}$ said he hoped to one day become an [astronaut]$_{head:title}$	*per:title*	11
	Neither he nor [Aquash]$_{head:per}$ were [American]$_{tail:nationality}$ citizens .	*per:origin*	
Relation Def.	[Zhang Yinjun]$_{tail:per}$, spokesperson with one of China 's largest charity organization , the [China Charity Federation]$_{head:org}$	*org:top_mem.*	96
Context Ignored	[Bibi]$_{head:per}$, a mother of [five]$_{tail:number}$, was sentenced this month to death .	*per:age*	52
No Relation	[He]$_{head:per}$ turned a gun on himself committing [suicide]$_{tail:causeofdeath}$.	*no_relation*	646

- 1017 examples, categorized into 9 error categories
- 7 categories related to context

Results (2): further error categories

Arguments			
Span	This is a tragic day for the Australian [Defence Force]$_{head:org}$ ([ADF]$_{tail:org}$)	*org:alt._nam*	12
Entity Type	[Christopher Bollyn]$_{head:per}$ is an [independent]$_{tail:religion}$ journalist	*per:religion*	31
	The company, which [Baldino]$_{head:org}$ founded in [1987]$_{tail:date}$ sells a variety of drugs	*org:founded*	

- Context misinterpretations account for ~96% of errors
- Argument errors account for ~4% of errors
- Incorrect assignment of “no relation” is the most common error

Evaluation

Automated analysis

Approach:

- Extend misclassification categories to testable hypotheses (error groups)
  - Group examples according to attribute, e.g., “has distracting entity in context”
  - Automatically verifiable on whole dataset split
- Validate whether the hypothesis holds
  - I.e., whether a group of instances shows an above average error rate
  - Based on the approach of [Wu et al., 2019]

Evaluation

Examples

TACRED (106k examples, 41 relations)

1. Data selection

based on errors of 49 different RE models

2. Human evaluation

re-annotation of 5k examples (dev and test)

3. Misclassification annotation

9 error categories

4. Automated analysis

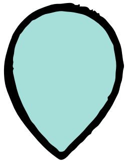
Evaluation

Examples — TACRED (106k examples, 41 relations)

1. Data selection — based on errors of 49 different RE models
2. Human evaluation — re-annotation of 5k examples (dev and test)
3. Misclassification annotation — 9 error categories
4. Automated analysis — 10 error groups

Evaluation

Error groups



Surface structure

e.g., argument
distance or
sentence length

Evaluation

Error groups

Surface structure

e.g., argument distance or sentence length

Arguments

e.g., head and tail entity type

Evaluation

Error groups

Surface structure

e.g., argument distance or sentence length

Arguments

e.g., head and tail entity type

Context

e.g., distracting entities in context

Evaluation

Error groups

Surface structure

e.g., argument distance or sentence length

Arguments

e.g., head and tail entity type

Context

e.g., distracting entities in context

Ground truth

e.g., positive examples (excluding “no relation”)

Evaluation

Error groups

Surface structure

e.g., argument distance or sentence length

Arguments

e.g., head and tail entity type

Context

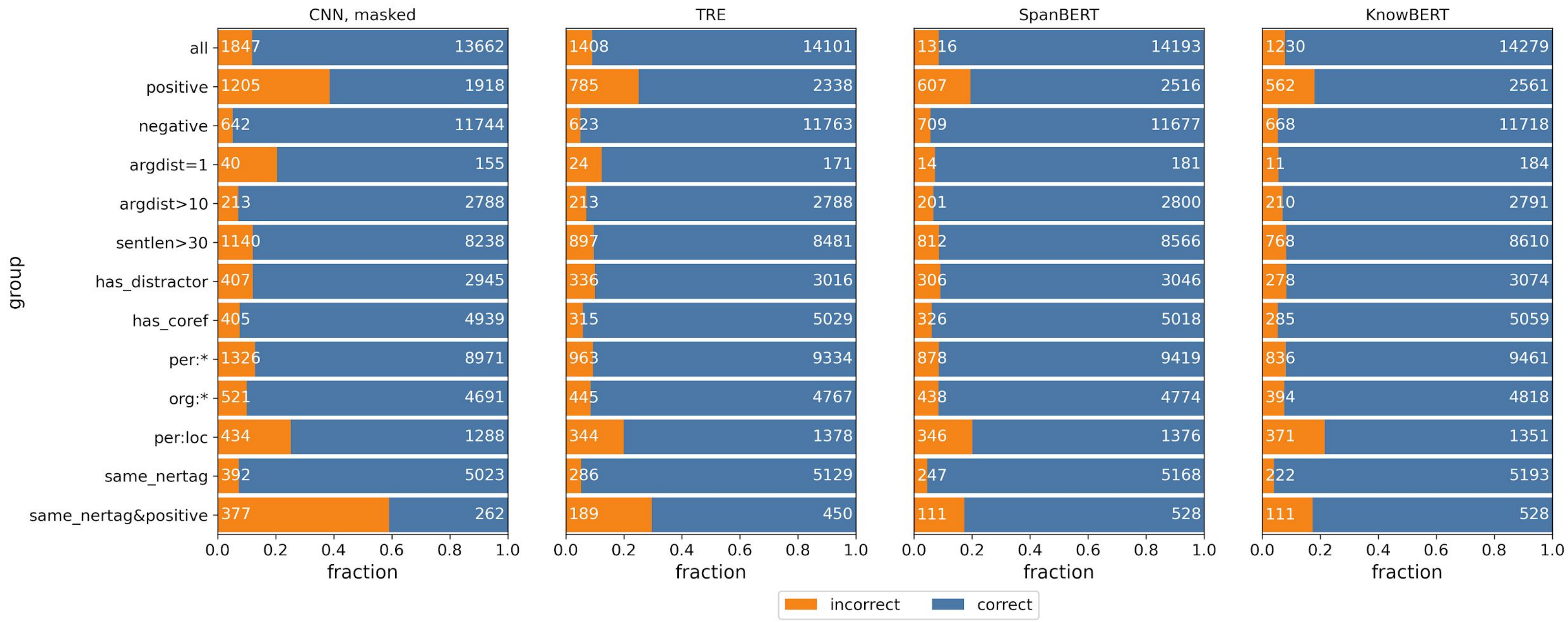
e.g., distracting entities in context

Ground truth

e.g., positive examples (excluding “no relation”)

- Compare state-of-the-art model error rates per group
  - TRE [Alt et al., 2019] → OpenAI GPT
  - SpanBERT [Joshi et al., 2019] → BERT, pre-trained on span level
  - KnowBERT [Peters et al., 2019] → BERT, pre-trained jointly with entity linking

Results (3): per group error rates



Results (3): per group error rates

- Large fraction of errors caused by two ambiguous groups of relations
  - per:loc relations expressed in similar context
    - e.g., *per:cities_of_resid.* vs. *per:countries_of_resid.*
  - same_nertag&positive have same argument types
    - e.g., *per:parents*, *per:children* and *per:other_family*

Summary

- Manual re-annotation of 5k most challenging TACRED examples (development and test split)
  - → Results: Release of revised dataset

Lessons learned:

- Careful evaluation of development and test splits necessary if dataset is crowdsourced
  - → to ensure progress can be measured accurately
- Models often unable to predict a relation even if clearly expressed
- Models frequently ignore argument roles or ignore sentential context
- Two groups of ambiguous relations mainly responsible for remaining errors

Takeaways and future directions

- A clear definition of the (practical) purpose of the task
  - e.g., IE for knowledge base construction vs. question answering
- Probing → causation, i.e., encoded information actually impacts prediction [Elazar et al. 2020]
- More detailed investigation of datasets and linguistic phenomena
  - e.g., context vs. entity mentions [Peng et al., 2020] or via challenge sets [Rosenman et al., 2020]
- Pre-training focused on semantic relations

Thank you!
Questions?

Github: github.com/DFKI-NLP

Website: christophalt.github.io

Credits: Some of the icons and graphics were created by Slidesgo, including icons from Flaticon.

References

- [Zhang et al., 2017] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. EMNLP, 2017.
- [Hendrickx et al., 2010] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Seaghdha, Sebastian Pado, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. SemEval, 2010.
- [Manning et al., 2014] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. ACL 2014 (System Demonstrations).
- [Peters et al., 2019] Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. EMNLP, 2019.
- [Wu et al., 2019] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. Errudite: Scalable, reproducible, and testable error analysis. ACL, 2019.
- [Joshi et al., 2019] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke S. Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. TACL, 2019.
- [Alt et al., 2019] Christoph Alt*, Marc Hübner*, and Leonhard Hennig. Improving relation extraction by pre-trained language representations. AKBC, 2019.
- [Rosenman et al., 2020] Exposing Shallow Heuristics of Relation Extraction Models with Challenge Data. Shachar Rosenman, Alon Jacovi, Yoav Goldberg EMNLP 2020.
- [Peng et al., 2020] Learning from Context or Names? An Empirical Study on Neural Relation Extraction. Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, Jie Zhou. EMNLP 2020.
- [Elazar et al., 2020] When Bert Forgets How To POS: Amnesic Probing of Linguistic Properties and MLM Predictions. Yanai Elazar, Shauli Ravfogel, Alon Jacovi, Yoav Goldberg. arXiv 2020.