Neural sequential transfer learning for relation extraction

Christoph Alt

November 30, 2020

Chair: Prof. Dr. Klaus Obermayer
Supervisor: Prof. Dr.-Ing. Sebastian Möller
Reviewer: Prof. Dr. Hans Uszkoreit
Prof. Dr.-Ing. Alan Akbik

Outline

- Motivation & background
- Problem statement
- Objectives and contributions
- Sequential transfer learning for neural relation extraction
  - Approach
  - Evaluation
  - Experiments
- Conclusion
- Outlook

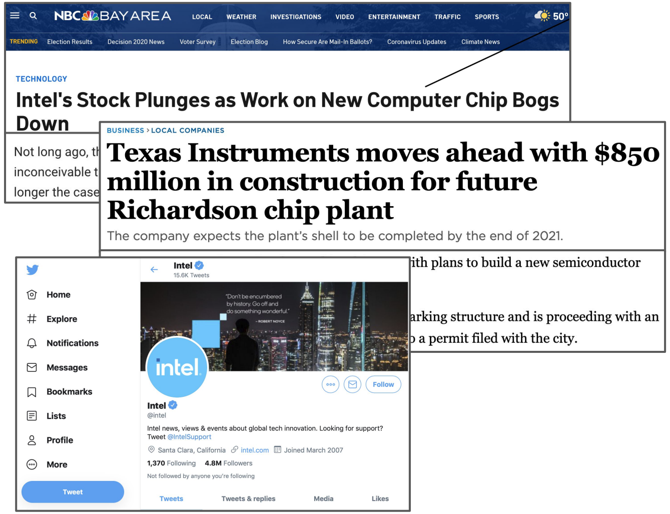
Information extraction

Information extraction

Information extraction

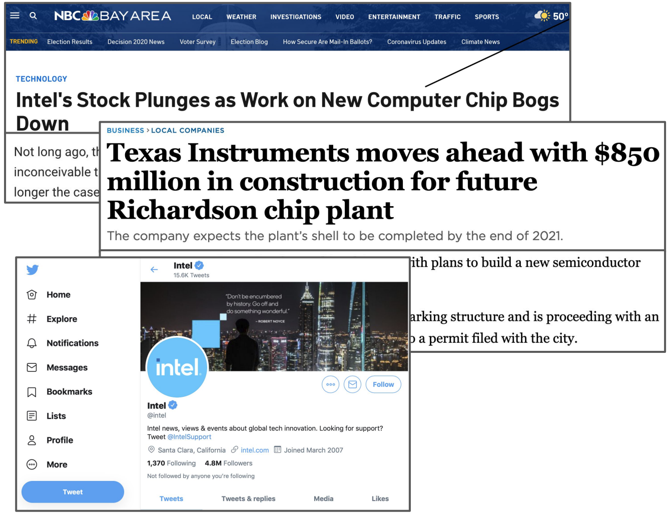
company	located in	...
Intel	Santa Clara	...
Texas Instruments	Dallas	...
...	...	...

Information extraction



company	located in	...
Intel	Santa Clara	...
Texas Instruments	Dallas	...
...	...	...

Information extraction



(Texas Instruments, located in, Dallas)

(Intel, located in, Santa Clara)

company	located in	…
Intel	Santa Clara	…
Texas Instruments	Dallas	…
…	…	…

Relation extraction

Relation extraction

- detect and retrieve relational information from unstructured text

Relation extraction

- detect and retrieve relational information from unstructured text

Intel is based in Santa Clara .

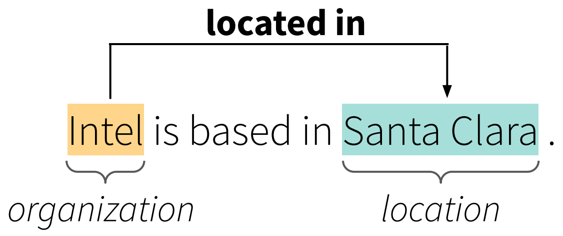
Relation extraction

- detect and retrieve relational information from unstructured text

$$\underbrace{\text{Intel}}_{\textit{organization}} \text{ is based in } \underbrace{\text{Santa Clara}}_{\textit{location}} \text{ .}$$

Relation extraction

- detect and retrieve relational information from unstructured text



Relation extraction

- detect and retrieve relational information from unstructured text

located in

Intel is based in Santa Clara .

organization *location*

vs.

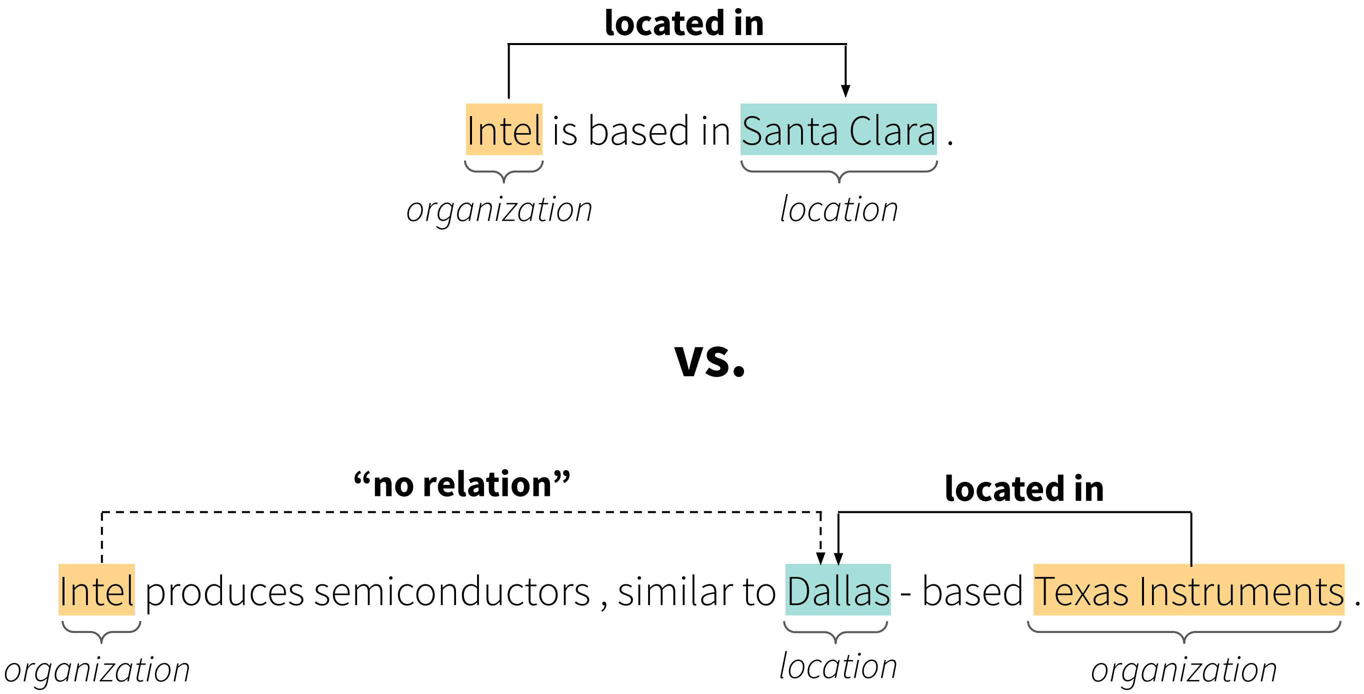
located in

Intel produces semiconductors , similar to Dallas - based Texas Instruments .

organization *location* *organization*

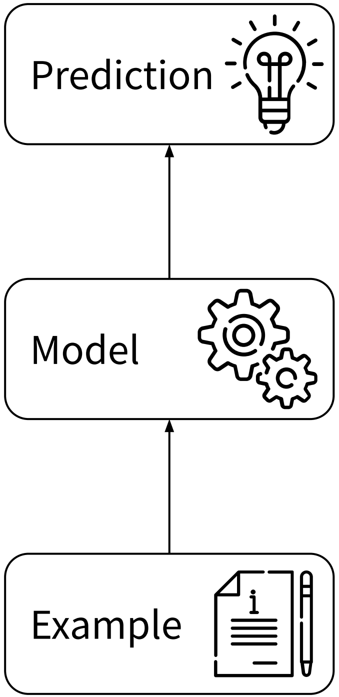
Relation extraction

- detect and retrieve relational information from unstructured text

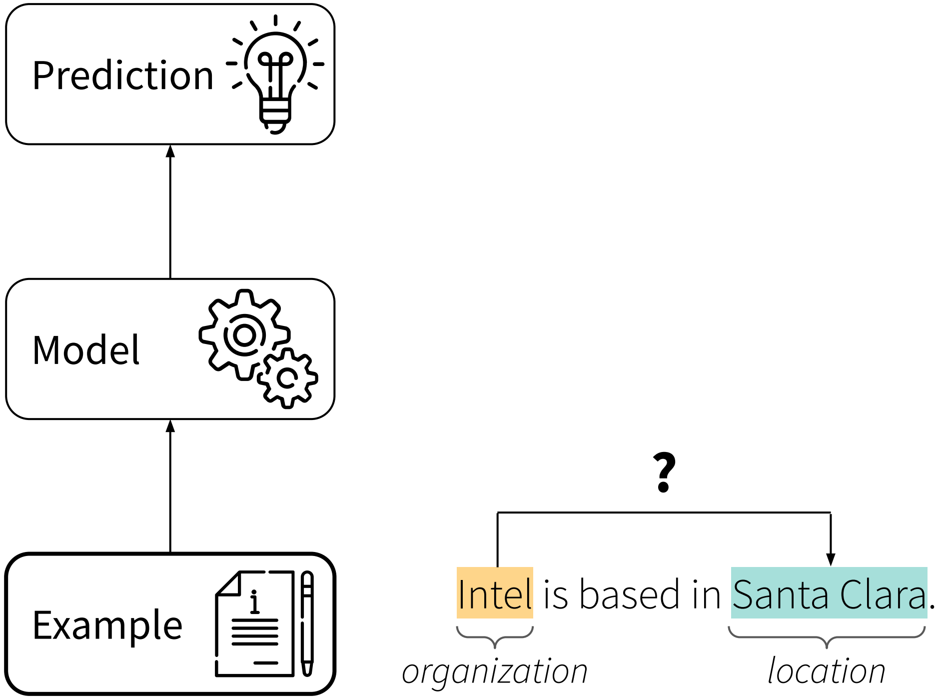


Machine-learning-based relation extraction

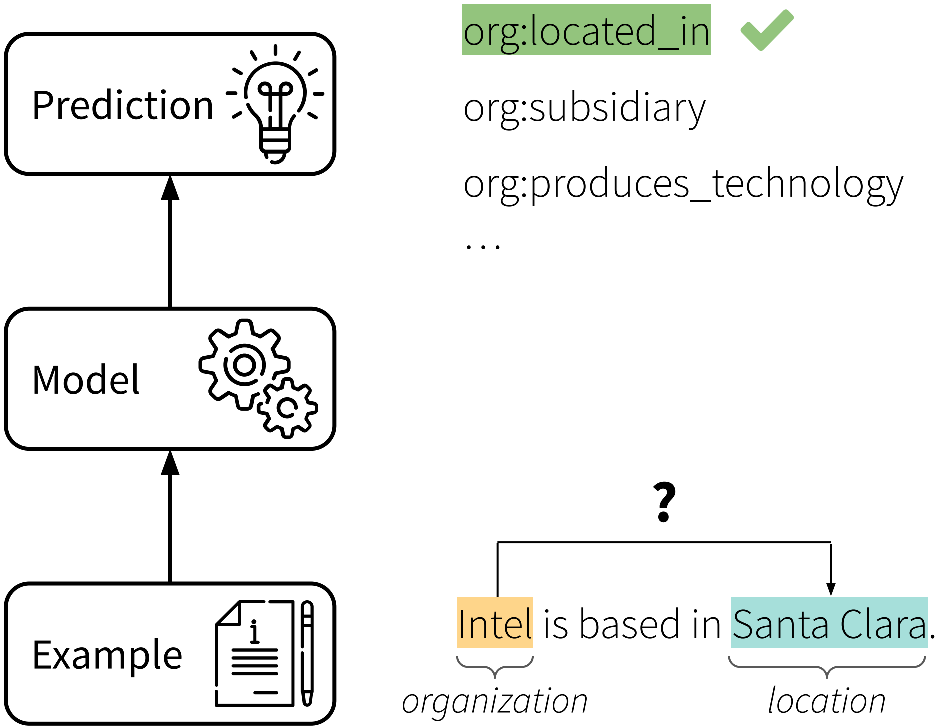
Machine-learning-based relation extraction



Machine-learning-based relation extraction



Machine-learning-based relation extraction



Relation extraction: Problems

- Quality and accuracy of extracted relations critical
- Neural-network-based methods achieve state-of-the-art results
  - problem: they are data-intensive

Relation extraction: Problems

- Quality and accuracy of extracted relations critical
- Neural-network-based methods achieve state-of-the-art results
  - problem: they are data-intensive

In practical scenarios

- Limited amount of supervised (labeled) data
- Model creation solely from task-specific data

Relation extraction: Problems

- Quality and accuracy of extracted relations critical
- Neural-network-based methods achieve state-of-the-art results
  - problem: they are data-intensive

In practical scenarios

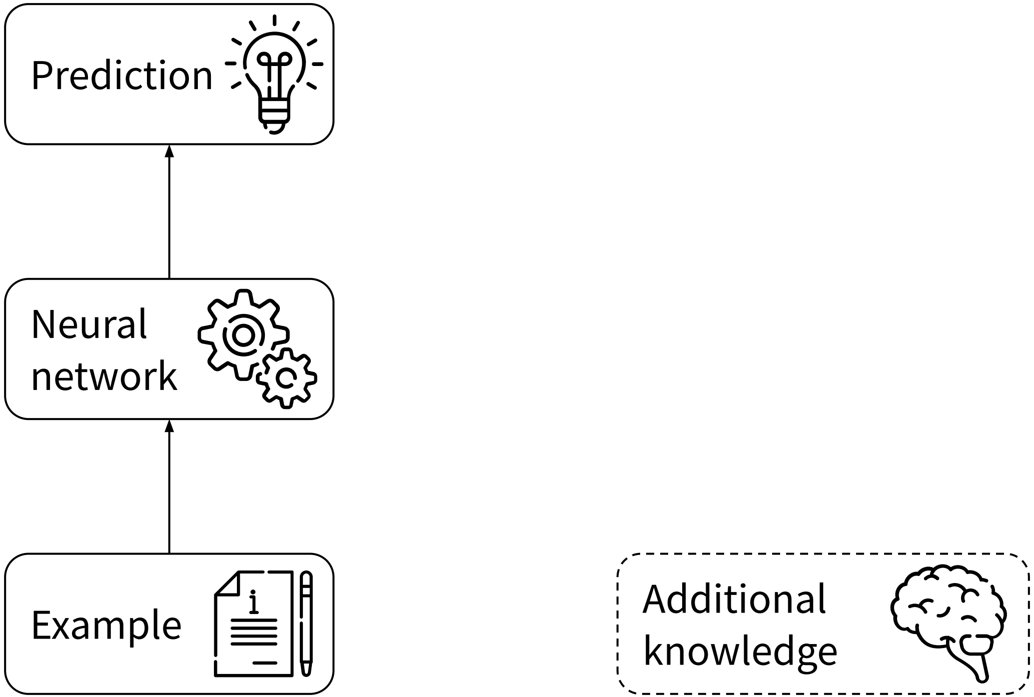
- Limited amount of supervised (labeled) data
- Model creation solely from task-specific data

Issue

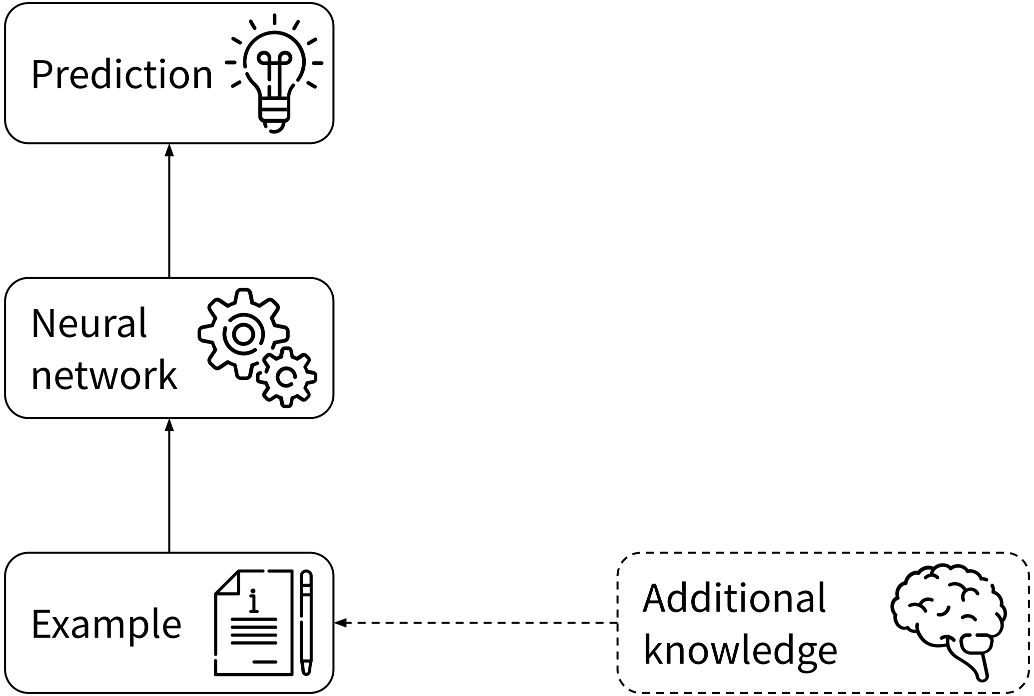
- Insufficient data to reliably model robust patterns

→ Poor generalization

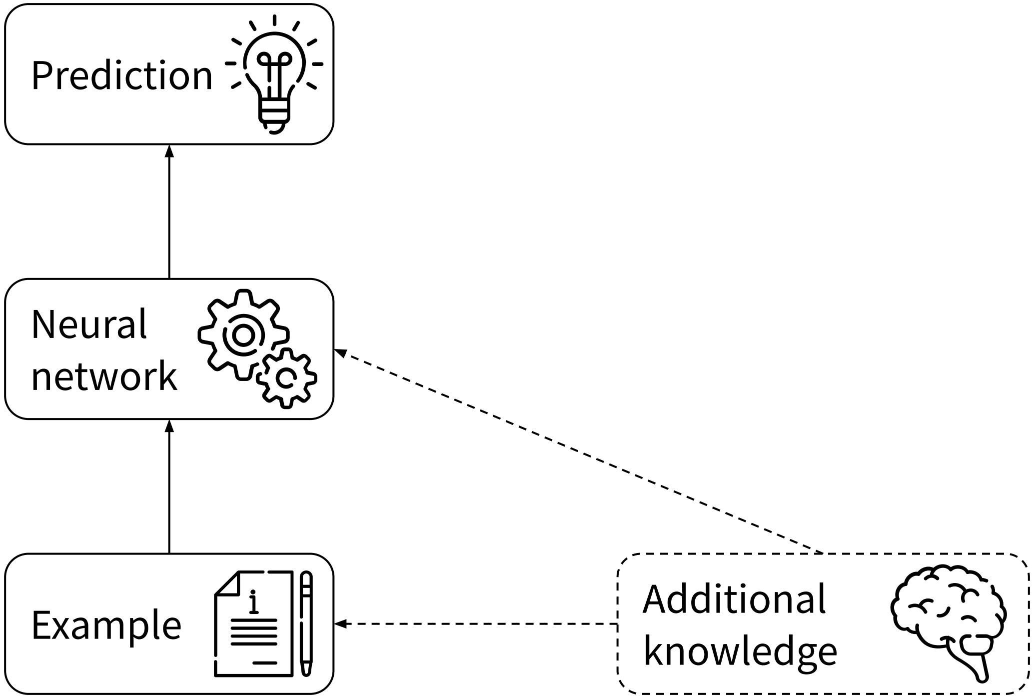
State-of-the-art: Improve generalization with explicit features



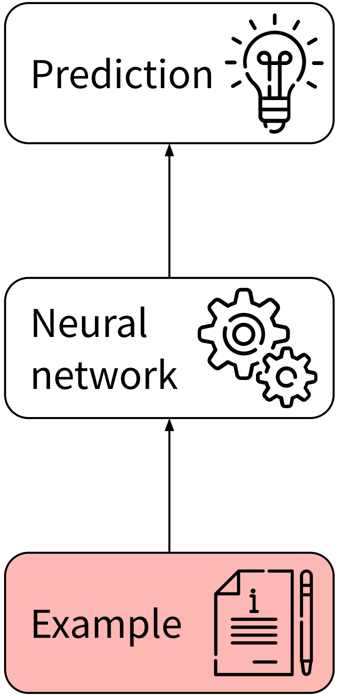
State-of-the-art: Improve generalization with explicit features



State-of-the-art: Improve generalization with explicit features



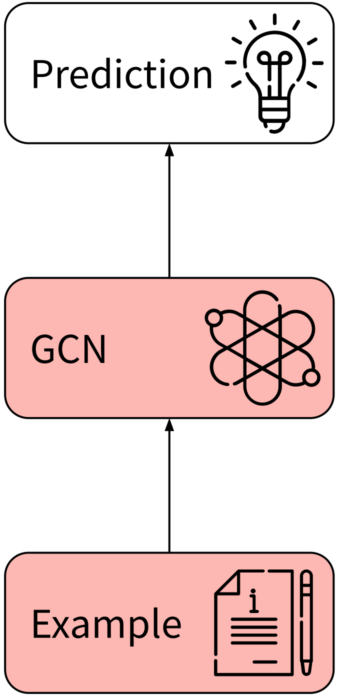
State-of-the-art: Improve generalization with explicit features

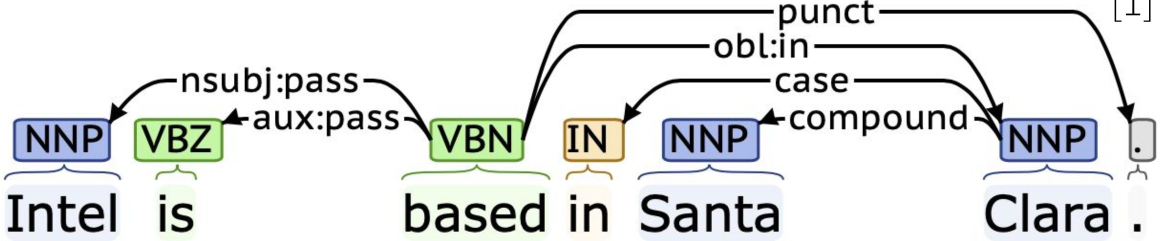


[1]

[1] visualization created with CoreNLP 4.0.0 [Manning et al., 2014]

State-of-the-art: Improve generalization with explicit features





[1] visualization created with CoreNLP 4.0.0 [Manning et al., 2014]

State-of-the-art: Challenges

State-of-the-art: Challenges

- **Complexity:** multiple systems (feature extractors), task-specific model architecture

State-of-the-art: Challenges

- **Complexity:** multiple systems (feature extractors), task-specific model architecture
- **Error propagation:** errors can propagate and accumulate

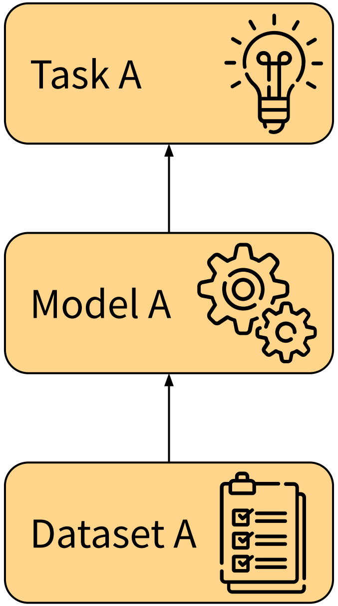
State-of-the-art: Challenges

- **Complexity:** multiple systems (feature extractors), task-specific model architecture
- **Error propagation:** errors can propagate and accumulate
- **Limited portability:** domain and language dependence

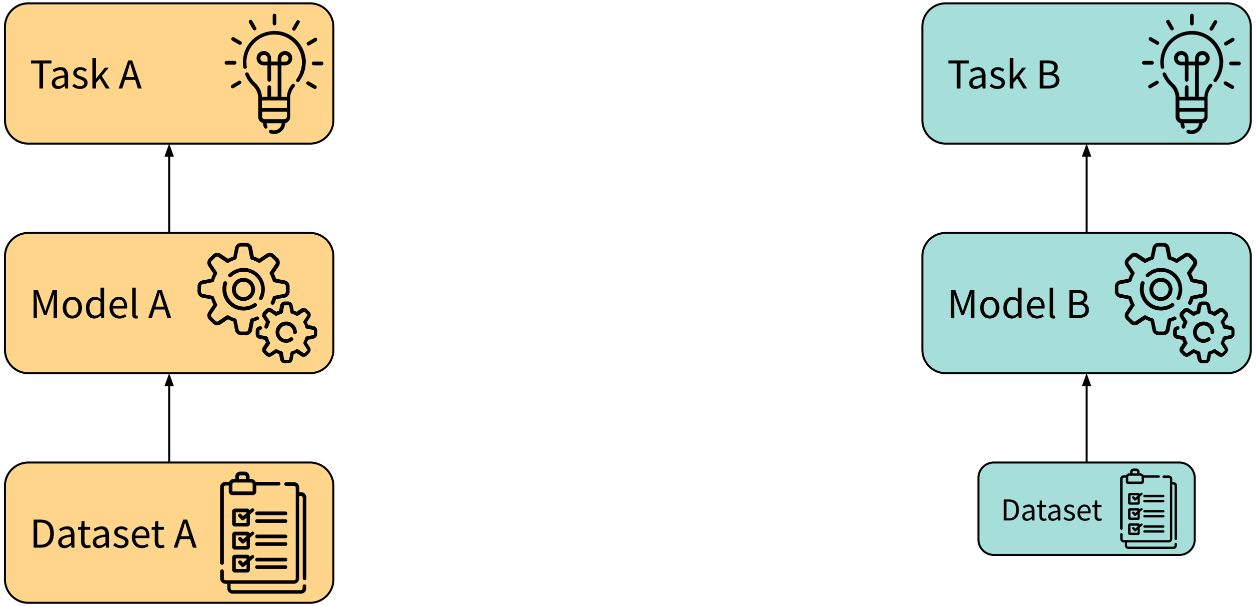
State-of-the-art: Challenges

- **Complexity:** multiple systems (feature extractors), task-specific model architecture
- **Error propagation:** errors can propagate and accumulate
- **Limited portability:** domain and language dependence
- **A-priori feature selection:** features selected before training

Transfer learning

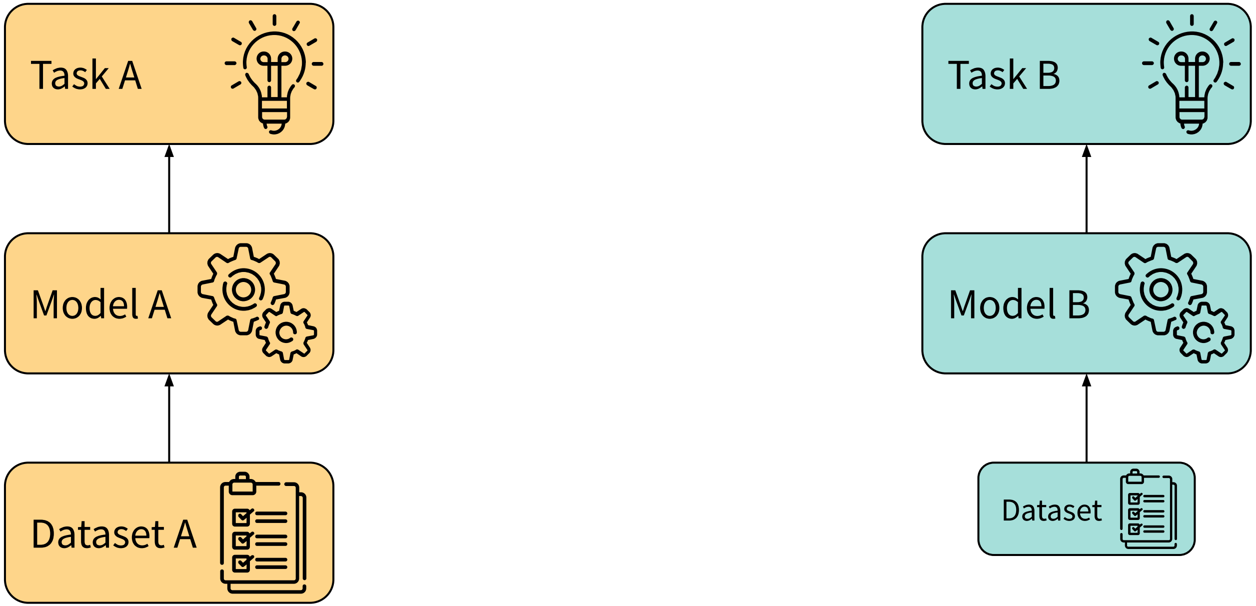


Transfer learning

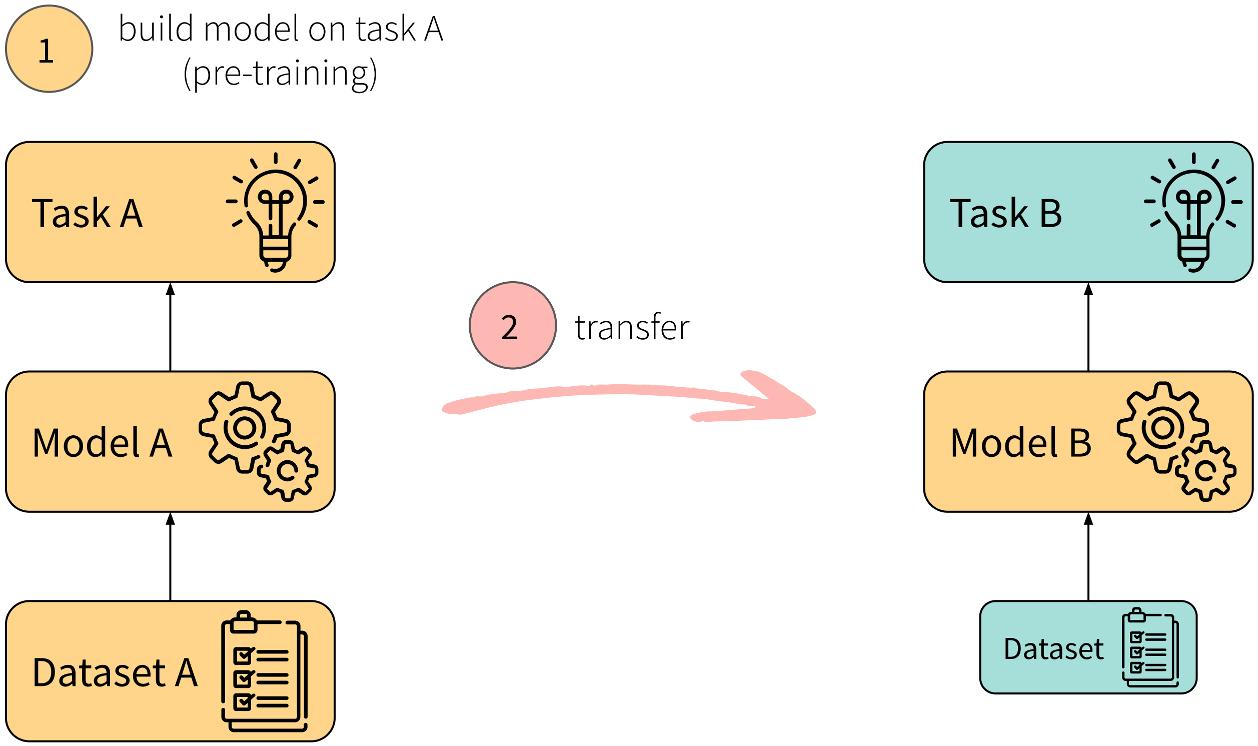


Transfer learning

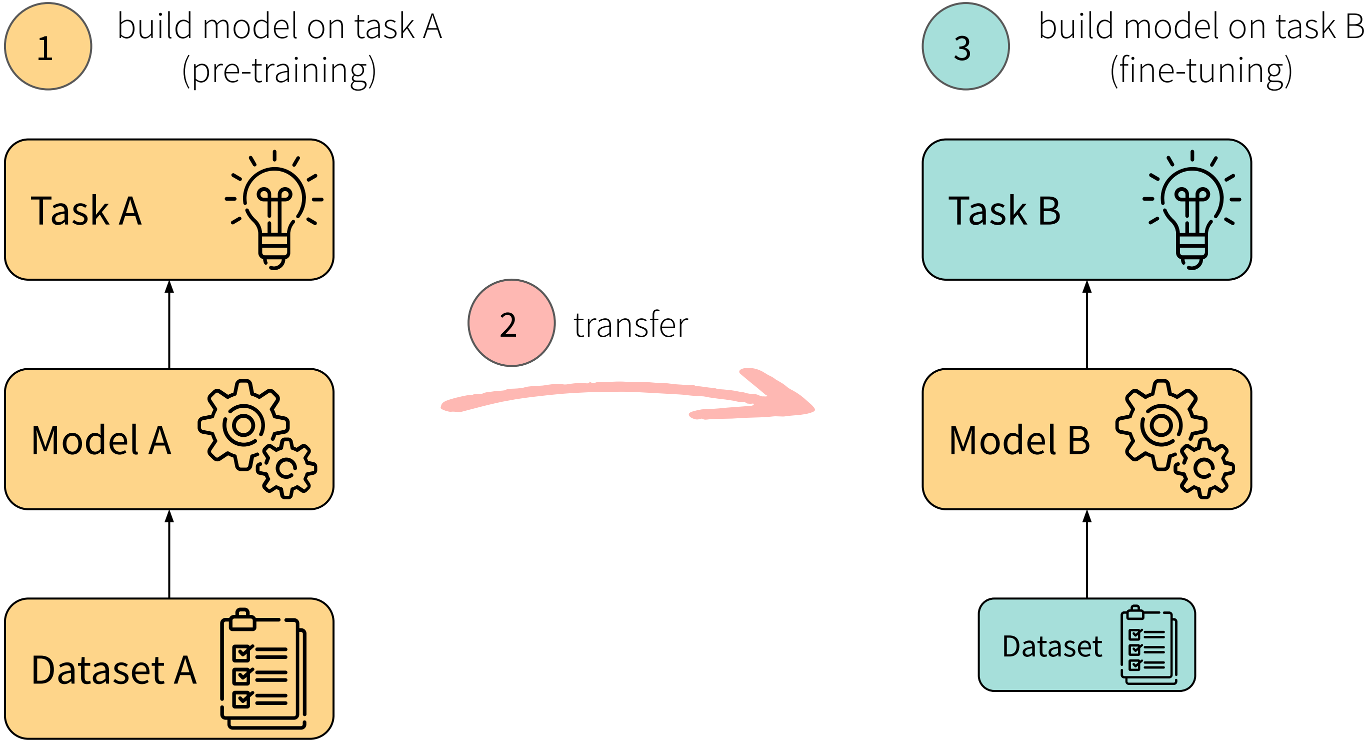
1 build model on task A
(pre-training)



Transfer learning



Transfer learning



Objectives and contributions

Develop better performing and more data-efficient neural relation extraction methods

Objectives and contributions

Develop better performing and more data-efficient neural relation extraction methods

Main contributions

Sequential transfer learning for supervised relation extraction

C. Alt*, M. Hübner*, L. Hennig. *"Improving Relation Extraction by Pre-trained Language Representations"*. **AKBC 2019**.

Combining sequential transfer learning and distant supervision

C. Alt, M. Hübner, L. Hennig. *"Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction"*. **ACL 2019**.

Objectives and contributions

Develop better performing and more data-efficient neural relation extraction methods → Improve our understanding of neural relation extraction approaches

Main contributions

Sequential transfer learning for supervised relation extraction

C. Alt*, M. Hübner*, L. Hennig. *"Improving Relation Extraction by Pre-trained Language Representations"*. **AKBC 2019**.

Combining sequential transfer learning and distant supervision

C. Alt, M. Hübner, L. Hennig. *"Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction"*. **ACL 2019**.

Objectives and contributions

Develop better performing and more data-efficient neural relation extraction methods

→

Improve our understanding of neural relation extraction approaches

Main contributions

Sequential transfer learning for supervised relation extraction

C. Alt*, M. Hübner*, L. Hennig. *“Improving Relation Extraction by Pre-trained Language Representations”*. **AKBC 2019**.

Combining sequential transfer learning and distant supervision

C. Alt, M. Hübner, L. Hennig. *“Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction”*. **ACL 2019**.

Analyzing captured linguistic knowledge

C. Alt, A. Gabryszak, L. Hennig. *“Probing Linguistic Features of Sentence-Level Representations in Neural Relation Extraction”*. **ACL 2020**.

Fine-grained analysis of model errors and datasets

C. Alt, A. Gabryszak, L. Hennig. *“TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task”*. **ACL 2020**.

Objectives and contributions

Develop better performing and more data-efficient neural relation extraction methods

Improve our understanding of neural relation extraction approaches

Main contributions

Sequential transfer learning for supervised relation extraction

C. Alt*, M. Hübner*, L. Hennig. *"Improving Relation Extraction by Pre-trained Language Representations"*. **AKBC 2019**.

Combining sequential transfer learning and distant supervision

C. Alt, M. Hübner, L. Hennig. *"Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction"*. **ACL 2019**.

Analyzing captured linguistic knowledge

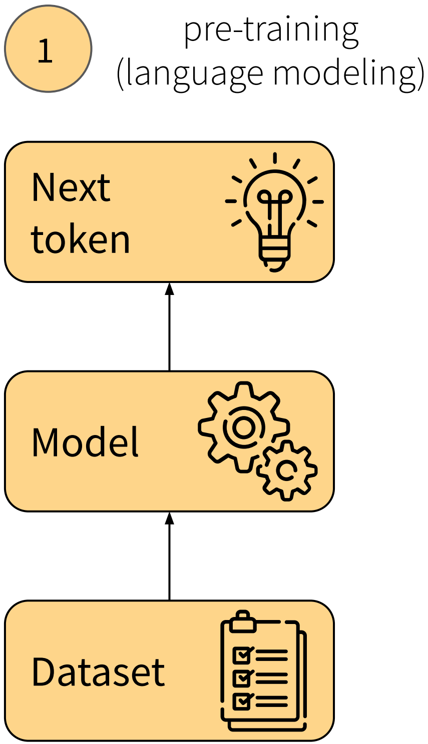
C. Alt, A. Gabryszak, L. Hennig. *"Probing Linguistic Features of Sentence-Level Representations in Neural Relation Extraction"*. **ACL 2020**.

Fine-grained analysis of model errors and datasets

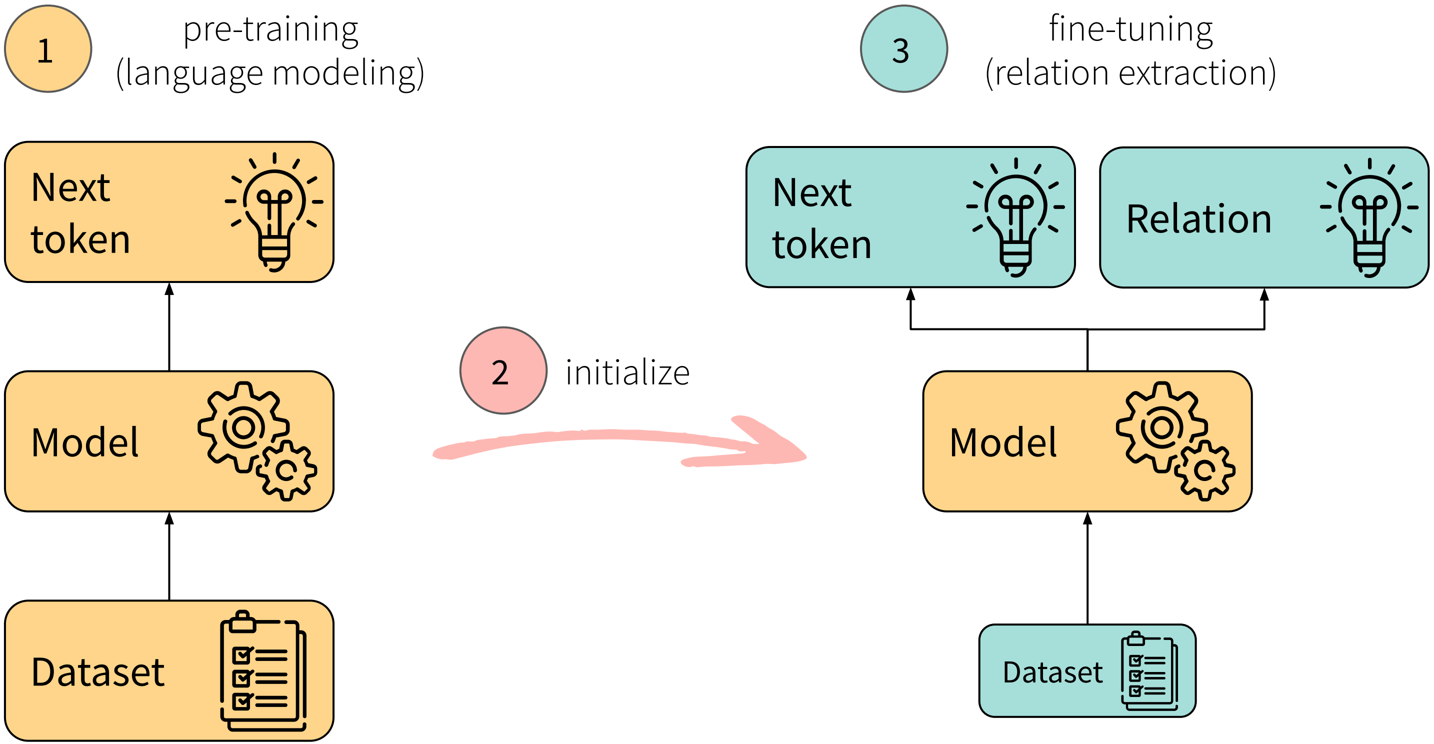
C. Alt, A. Gabryszak, L. Hennig. *"TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task"*. **ACL 2020**.

Algorithm

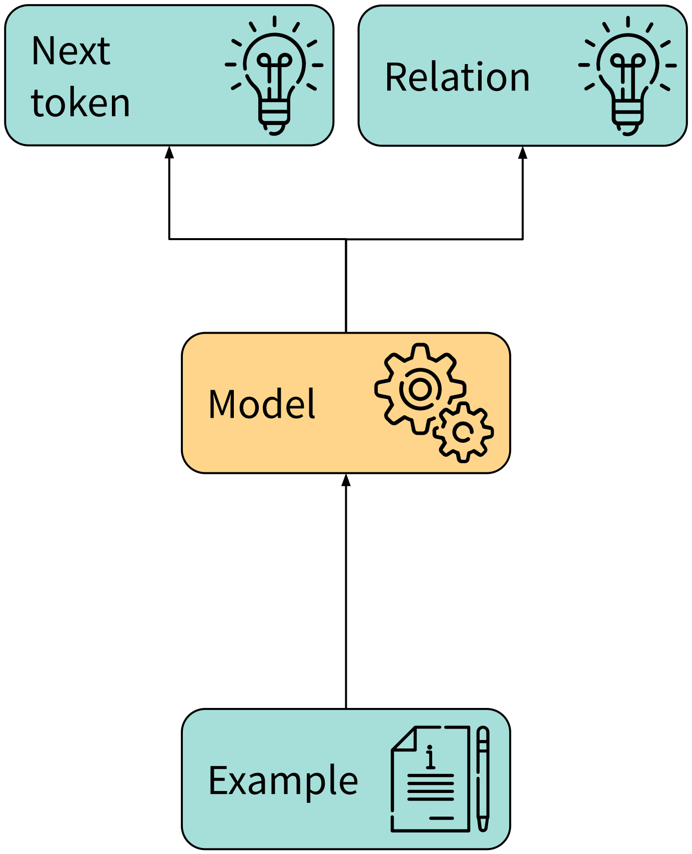
Algorithm



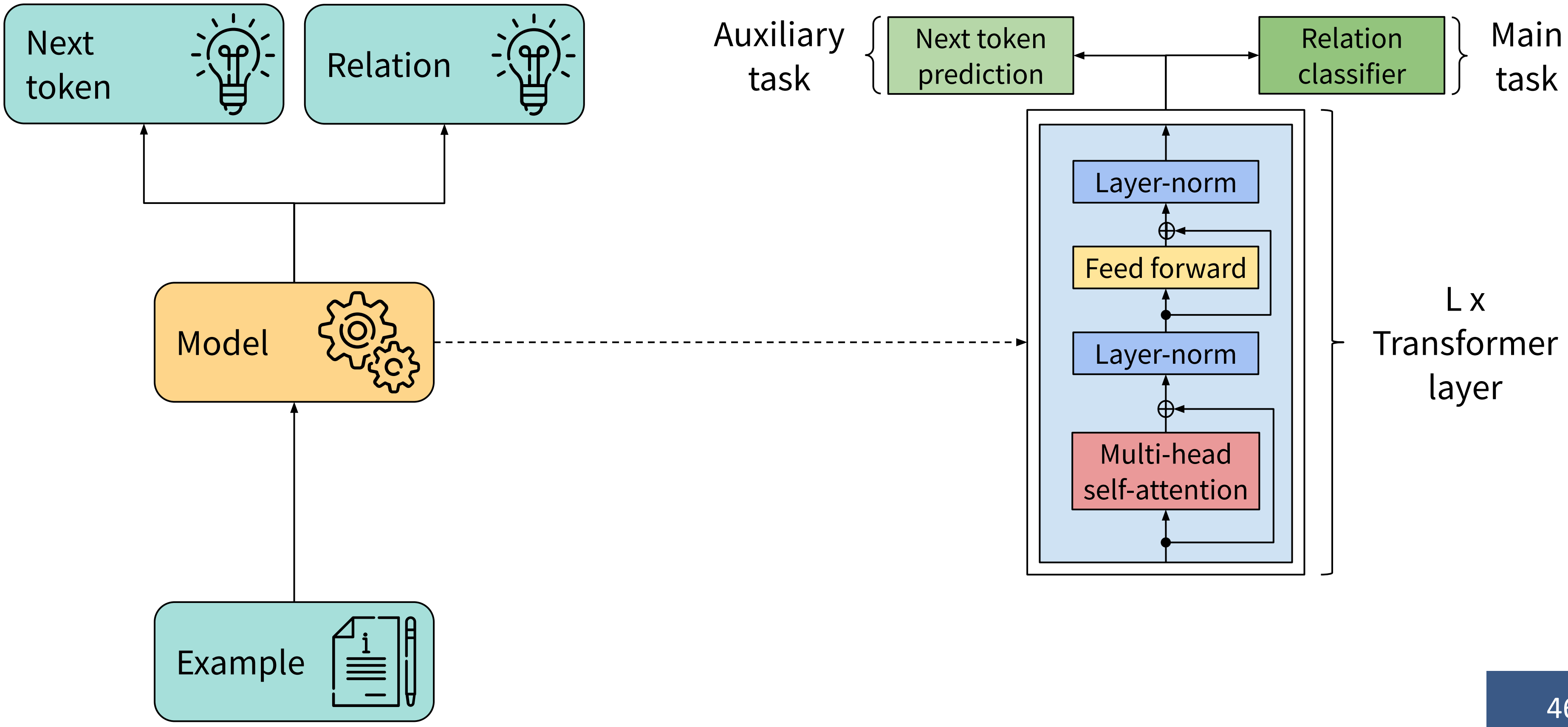
Algorithm


1
pre-training
(language modeling)
Next token
Model
Dataset
2
initialize
3
fine-tuning
(relation extraction)
Next token
Relation
Model
Dataset

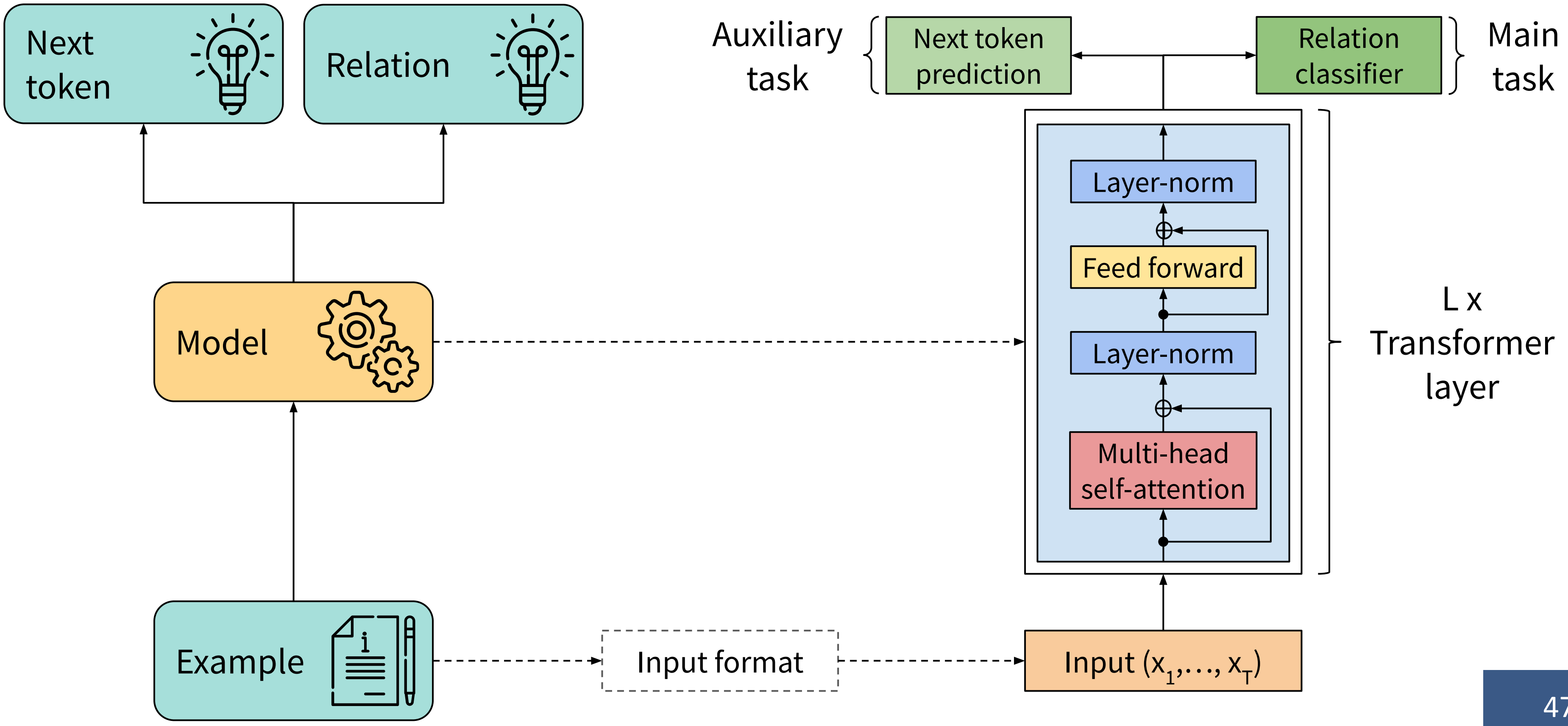
Model architecture



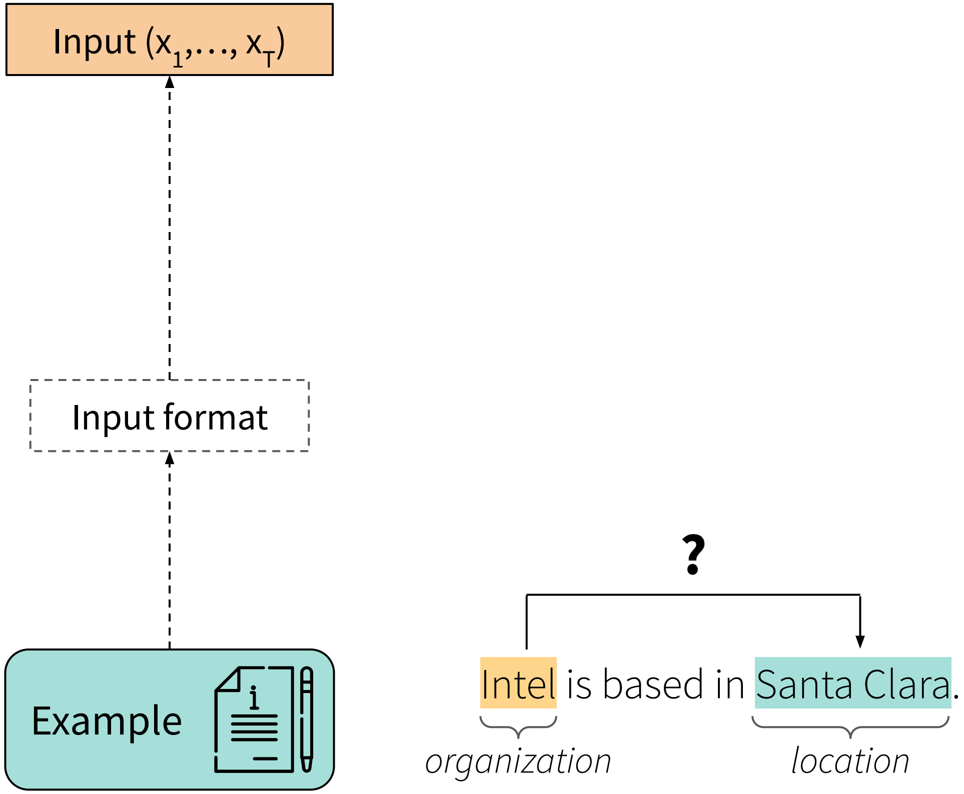
Model architecture



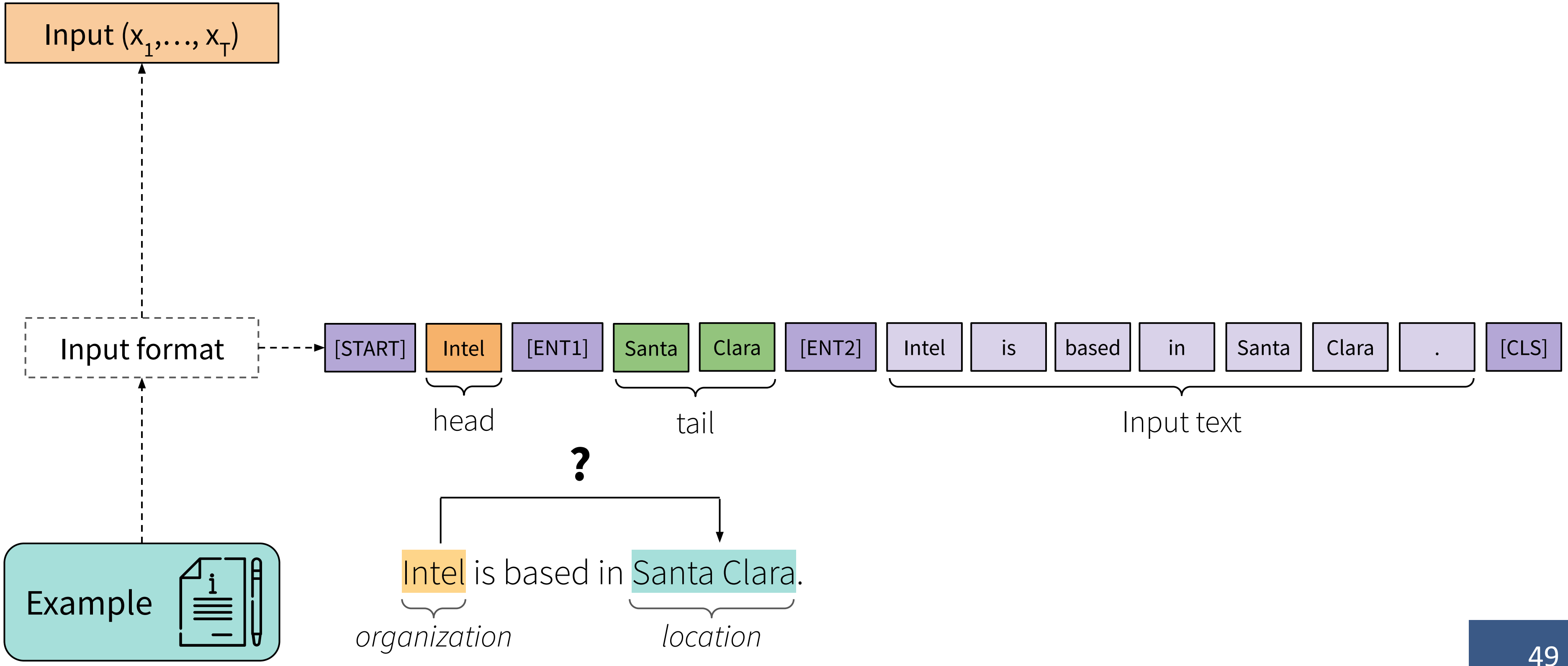
Model architecture



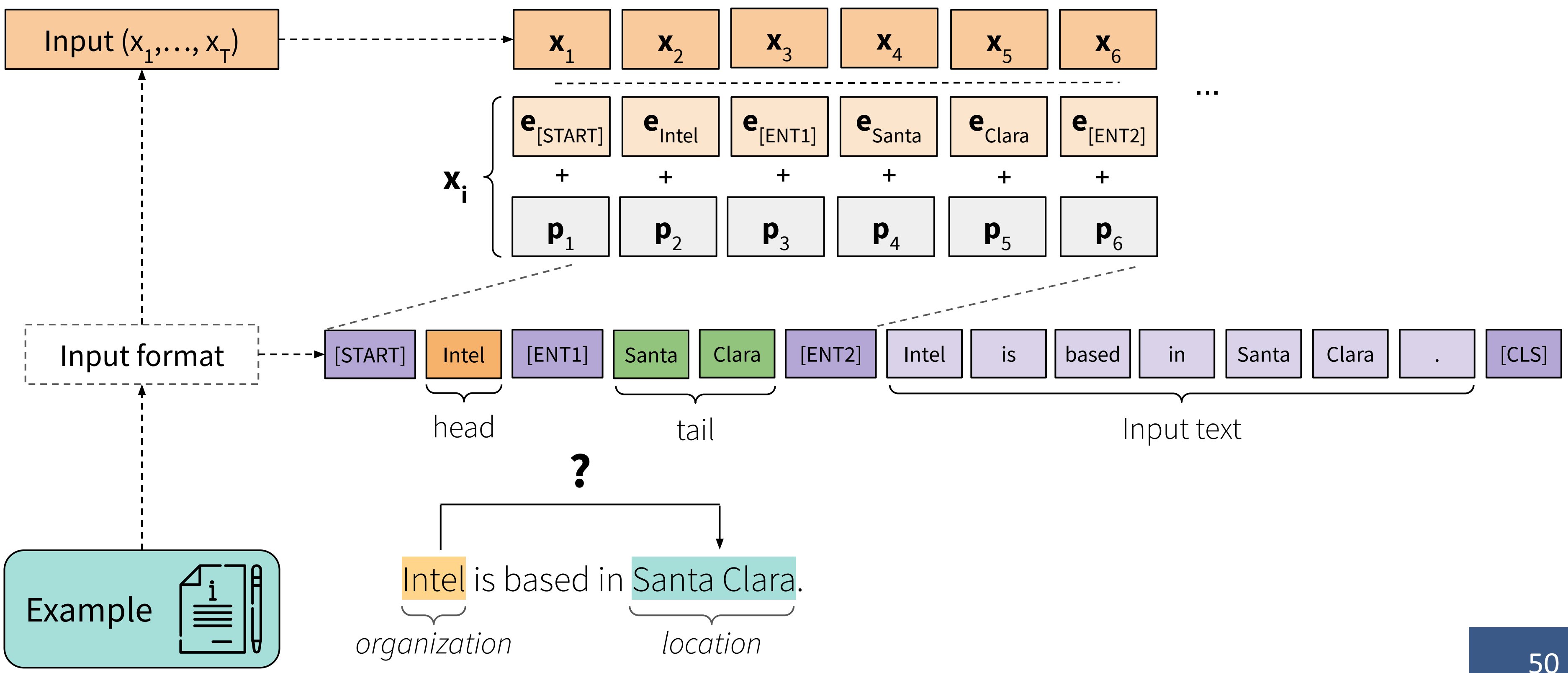
Input format



Input format

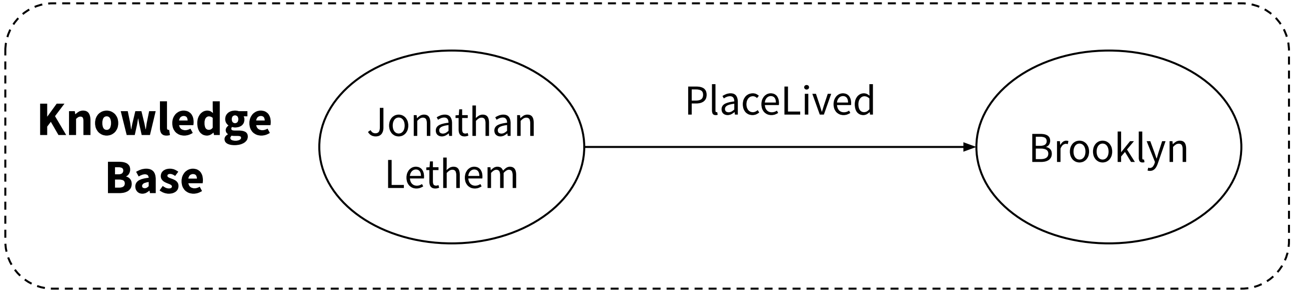


Input format



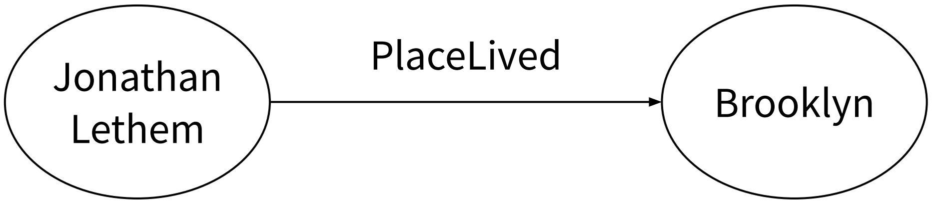
Distant supervision

Distant supervision



Distant supervision

Knowledge Base



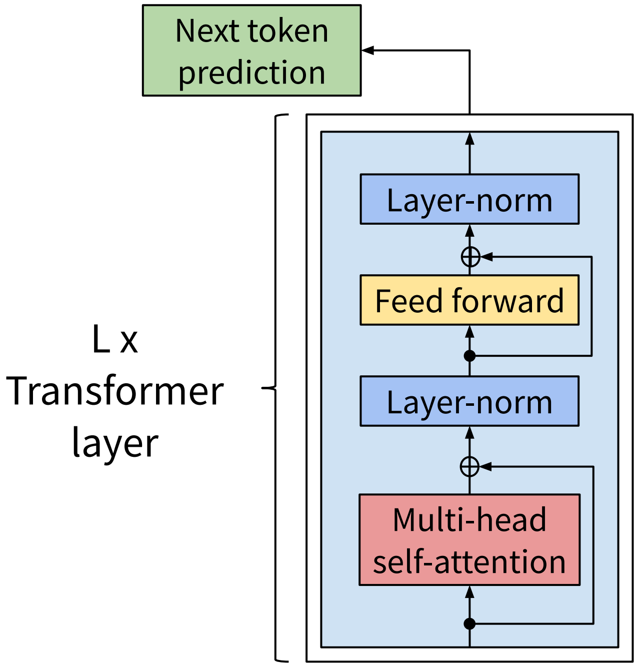
Data

You could say that only the dead, and Jonathan Lethem, know Brooklyn. ✗

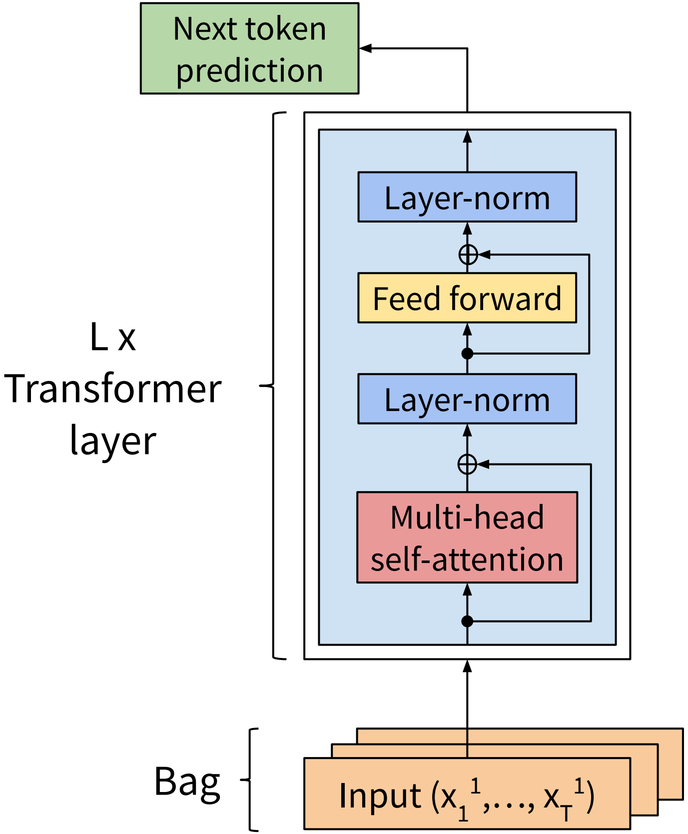
“Non-connectivity becomes a commodity , something to cherish, '' said Jonathan Lethem, a Brooklyn novelist and a new MacArthur fellow. ✓

In Brooklyn, they ask when you're going on Charlie Rose and if you know Jonathan Lethem. ✗

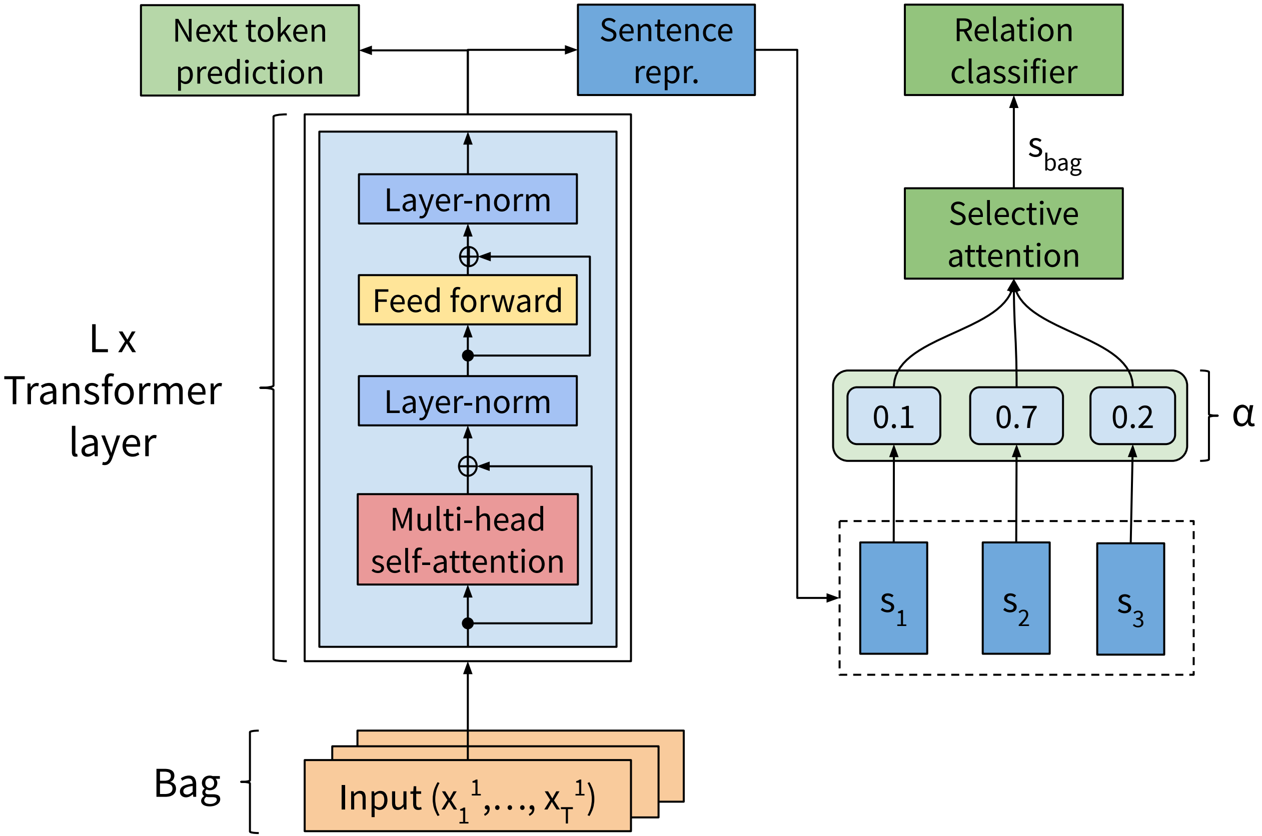
Extension to distantly supervised data



Extension to distantly supervised data



Extension to distantly supervised data



Parameter estimation

Parameter estimation

Relation extraction objective

$$L_{rel}(\mathcal{D}) = \sum_{i=1}^{|\mathcal{D}|} \log P(r_i | t_i^1, \ldots, t_i^{|T_i|}, head_i, tail_i)$$

Parameter estimation

Relation extraction objective

$$L_{rel}(\mathcal{D}) = \sum_{i=1}^{|\mathcal{D}|} \log P(r_i | t_i^1, \ldots, t_i^{|T_i|}, head_i, tail_i) \rightarrow f_R(f_M(\ldots;\theta_M);\theta_R)$$

Parameter estimation

Relation extraction objective

$$L_{rel}(\mathcal{D}) = \sum_{i=1}^{|\mathcal{D}|} \log P(r_i | t_i^1, \ldots, t_i^{|T_i|}, head_i, tail_i)$$

$$f_R(f_M(\ldots; \theta_M); \theta_R)$$

Language model objective

$$L_{lang}(\mathcal{D}) = \sum_{i=1}^{|D|} \sum_{j=1}^{|T_i|} \log P(t_j | t_{j-1}, \ldots, t_1)$$

Parameter estimation

Relation extraction objective

$$L_{rel}(\mathcal{D}) = \sum_{i=1}^{|\mathcal{D}|} \log P(r_i | t_i^1, \ldots, t_i^{|T_i|}, head_i, tail_i) \rightarrow f_R(f_M(\ldots;\theta_M);\theta_R)$$

Language model objective

$$L_{lang}(\mathcal{D}) = \sum_{i=1}^{|D|} \sum_{j=1}^{|T_i|} \log P(t_j | t_{j-1}, \ldots, t_1) \rightarrow f_L(f_M(\ldots;\theta_M);\theta_L)$$

Parameter estimation

Relation extraction objective

$$L_{rel}(\mathcal{D}) = \sum_{i=1}^{|\mathcal{D}|} \log P(r_i | t_i^1, \ldots, t_i^{|T_i|}, head_i, tail_i) \rightarrow f_R(f_M(\ldots;\theta_M);\theta_R)$$

Language model objective

$$L_{lang}(\mathcal{D}) = \sum_{i=1}^{|D|} \sum_{j=1}^{|T_i|} \log P(t_j | t_{j-1}, \ldots, t_1) \rightarrow f_L(f_M(\ldots;\theta_M);\theta_L)$$

Maximum likelihood estimate

$$L(\mathcal{D}) = L_{rel}(\mathcal{D}) + \lambda * L_{lang}(\mathcal{D})$$

$$\hat{\theta} = \arg\max_{\theta} L(\mathcal{D}; \theta), with\ \theta = \{\theta_M, \theta_R, \theta_L\}$$

Datasets

Dataset	Examples	Neg. examples (%)	Relations	Supervision
SemEval 2010 Task 8	10,717	17.4%	19	traditional
TACRED	106,264	79.5%	42	traditional
NYT-10	522,611	-	53	distant

Examples

SemEval 2010

content-container

The key was in a chest.

TACRED

org:subsidiary

The measure included Aerolineas's domestic subsidiary, Austral.

Evaluation

Evaluation

Hypothesis:

The proposed method performs equal or better than baselines that rely on explicit features.

Evaluation

Hypothesis:

The proposed method performs equal or better than baselines that rely on explicit features.

Experiment setup:

- Initialize the model (with parameters from OpenAI GPT [Radford et al., 2018])
- Fine-tune on the respective dataset
- Evaluate overall performance and data efficiency

Evaluation

Hypothesis:

The proposed method performs equal or better than baselines that rely on explicit features.

Experiment setup:

- Initialize the model (with parameters from OpenAI GPT [Radford et al., 2018])
- Fine-tune on the respective dataset
- Evaluate overall performance and data efficiency

Metrics:

- Performance: Precision, Recall, F1 score, P-R curve, area under the curve
- Data efficiency: F1 score over percentage of training data

Supervised RE: Results

TACRED

System	P	R	F1
LR	72.0	47.8	57.5
CNN	72.1	50.3	59.2
PCNN	**73.6**	53.4	61.9
Tree-LSTM	66.0	59.2	62.4
PA-LSTM	65.7	64.5	65.1
C-GCN	69.9	63.3	66.4
TRE	70.1	**65.0**	**67.4**

SemEval 2010

System	P	R	F1
SVM	–	–	82.2
PA-LSTM	–	–	82.7
C-GCN	–	–	84.8
DRNN	–	–	86.1
BRCNN	–	–	86.3
PCNN	86.7	86.7	86.6
TRE	88.0	86.2	**87.1**

Baselines: LR, SVM
State-of-the-art systems: PCNN, C-GCN, PA-LSTM

Supervised RE: Results

TACRED

System	P	R	F1
LR	72.0	47.8	57.5
CNN	72.1	50.3	59.2
PCNN	**73.6**	53.4	61.9
Tree-LSTM	66.0	59.2	62.4
PA-LSTM	65.7	64.5	65.1
C-GCN	69.9	63.3	66.4
TRE	70.1	**65.0**	**67.4**

SemEval 2010

System	P	R	F1
SVM	–	–	82.2
PA-LSTM	–	–	82.7
C-GCN	–	–	84.8
DRNN	–	–	86.1
BRCNN	–	–	86.3
PCNN	86.7	86.7	86.6
TRE	88.0	86.2	**87.1**

Baselines: LR, SVM
State-of-the-art systems: PCNN, C-GCN, PA-LSTM

Supervised RE: Results

TACRED

System	P	R	F1
LR	72.0	47.8	57.5
CNN	72.1	50.3	59.2
PCNN	**73.6**	53.4	61.9
Tree-LSTM	66.0	59.2	62.4
PA-LSTM	65.7	64.5	65.1
C-GCN	69.9	63.3	66.4
TRE	70.1	**65.0**	**67.4**

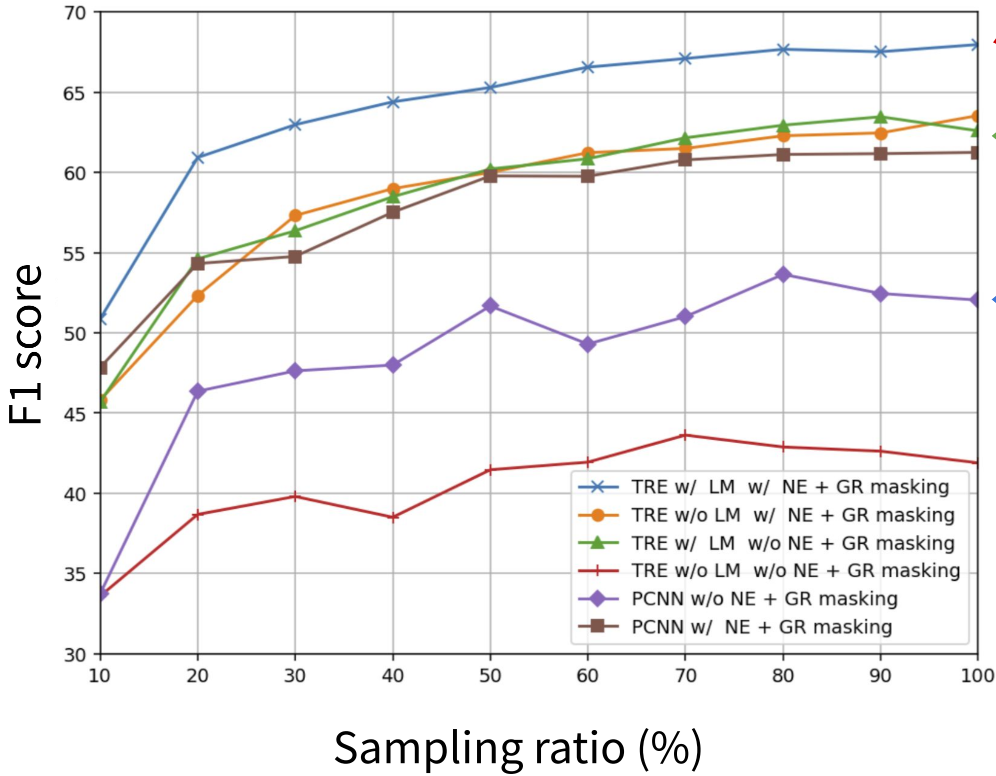
SemEval 2010

System	P	R	F1
SVM	–	–	82.2
PA-LSTM	–	–	82.7
C-GCN	–	–	84.8
DRNN	–	–	86.1
BRCNN	–	–	86.3
PCNN	86.7	86.7	86.6
TRE	88.0	86.2	**87.1**

Baselines: LR, SVM

State-of-the-art systems: PCNN, C-GCN, PA-LSTM

TACRED: Data efficiency

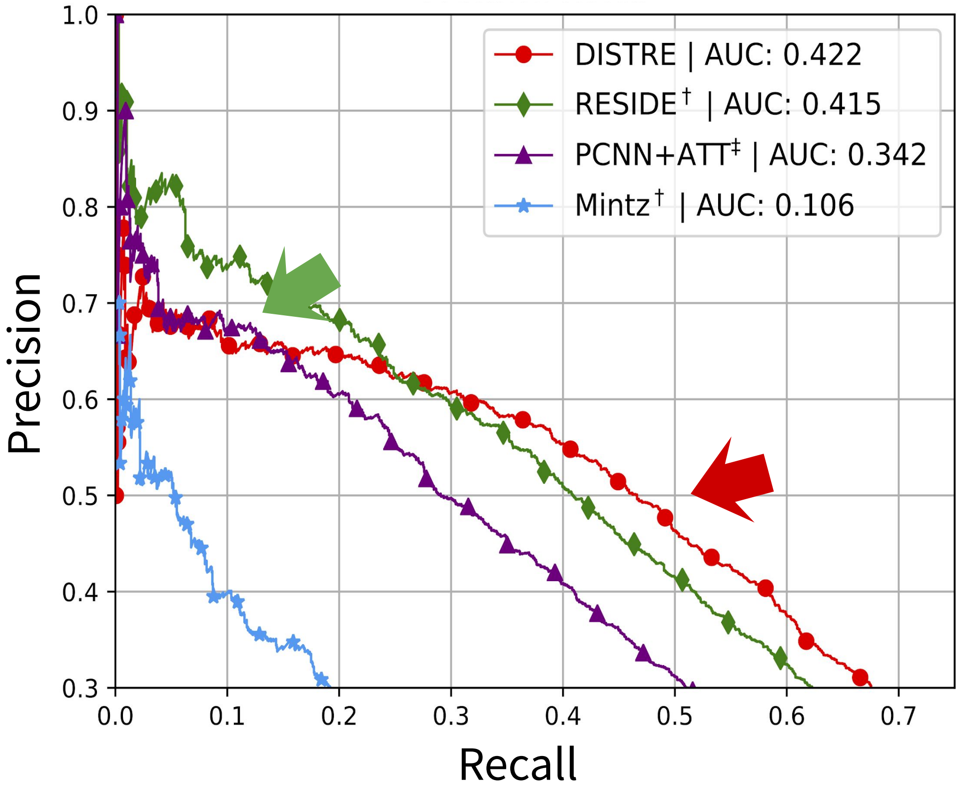


transfer learning with entity masking

transfer learning w/o entity masking
baseline with entity masking

baseline

Distantly supervised RE: Results



Baselines: Mintz

State-of-the-art system: RESIDE

Conclusion

Conclusion

- State-of-the-art sequential transfer learning systems for RE

Conclusion

- State-of-the-art sequential transfer learning systems for RE
- Language models capture more syntactic than semantic knowledge

Conclusion

- State-of-the-art sequential transfer learning systems for RE
- Language models capture more syntactic than semantic knowledge
- Improved performance on infrequently observed relations (long-tail)

Outlook

Outlook

- Improve acquisition and reuse of relevant knowledge

Outlook

- Improve acquisition and reuse of relevant knowledge
- Investigate other pre-training and multi-task learning strategies

Outlook

- Improve acquisition and reuse of relevant knowledge
- Investigate other pre-training and multi-task learning strategies
- Combine models for distantly supervised data

Outlook

- Improve acquisition and reuse of relevant knowledge
- Investigate other pre-training and multi-task learning strategies
- Combine models for distantly supervised data
- Further improvements require better understanding of models, datasets, and the task

Thank you!

Publications

- *Improving Relation Extraction by Pre-trained Language Representations.* Christoph Alt*, Marc Hübner* and Leonhard Hennig. **AKBC 2019**
- *Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction.* Christoph Alt, Marc Hübner and Leonhard Hennig. **ACL 2019**
- *Probing Linguistic Features of Sentence-Level Representations in Neural Relation Extraction.* Christoph Alt, Aleksandra Gabryszak and Leonhard Hennig. **ACL 2020**
- *TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task.* Christoph Alt, Aleksandra Gabryszak and Leonhard Hennig. **ACL 2020.**

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).
- [Radford et al., 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. arXiv 2018.