Schedule

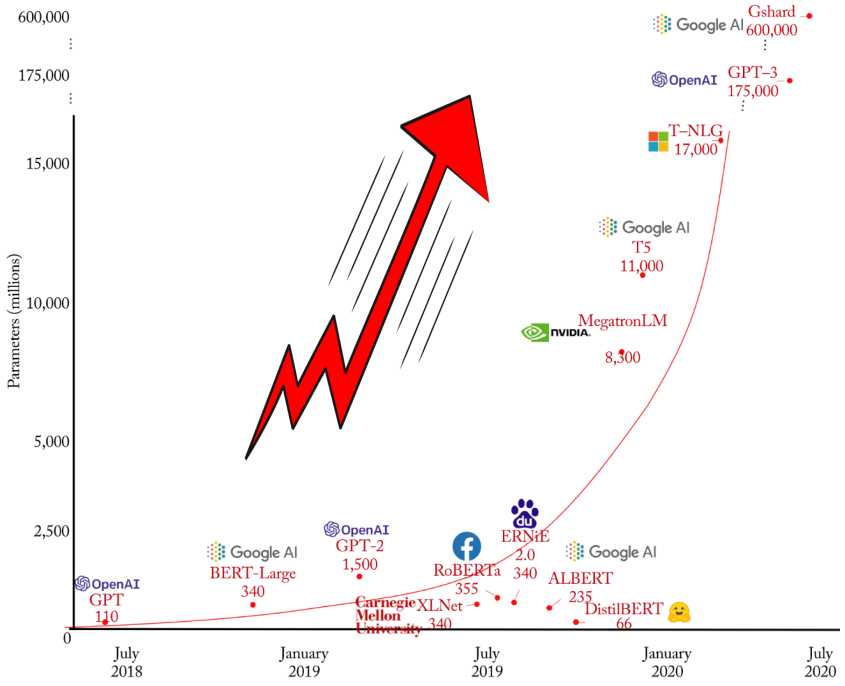
Part 3

Robust Knowledge Graph Construction

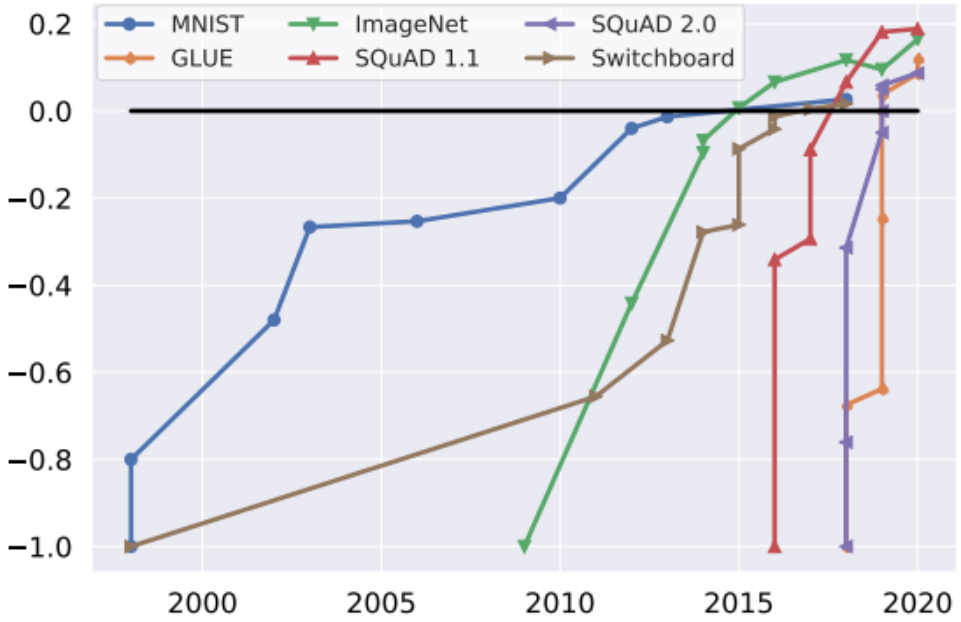
Tao Gui

Fudan University

Performance of Deep learning Models



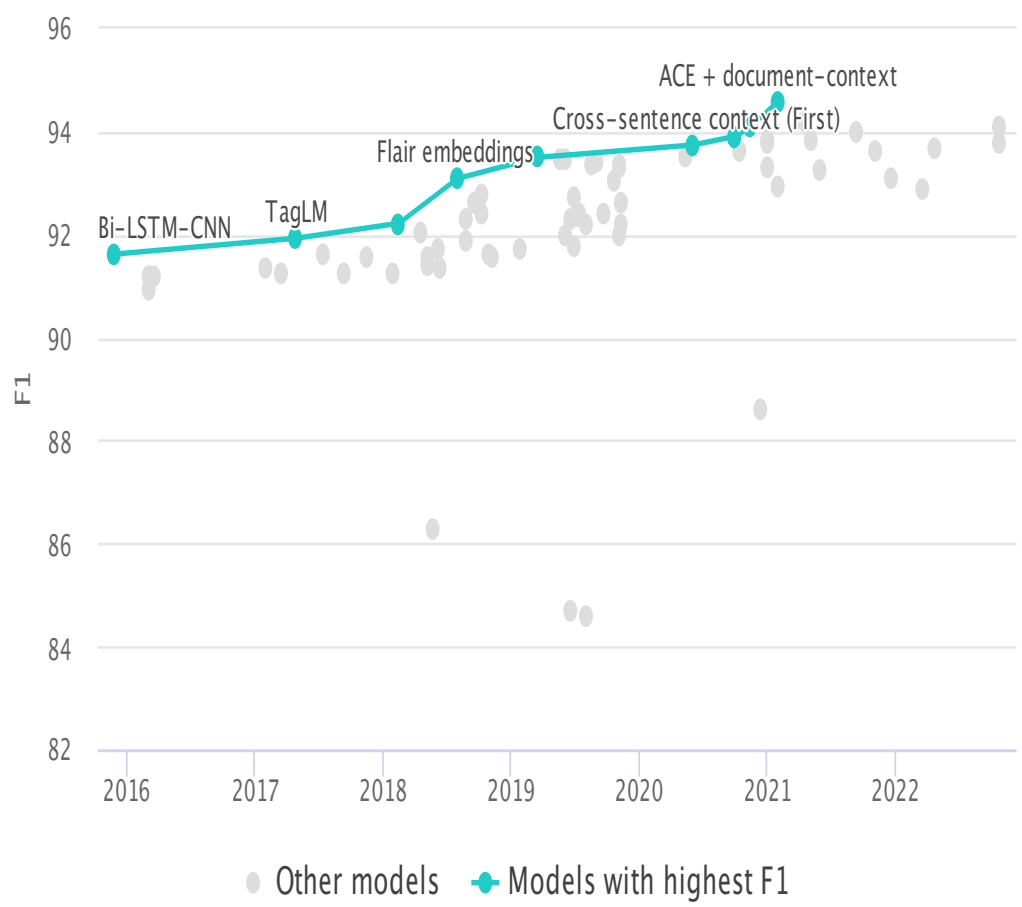
Number of Model Paremeters



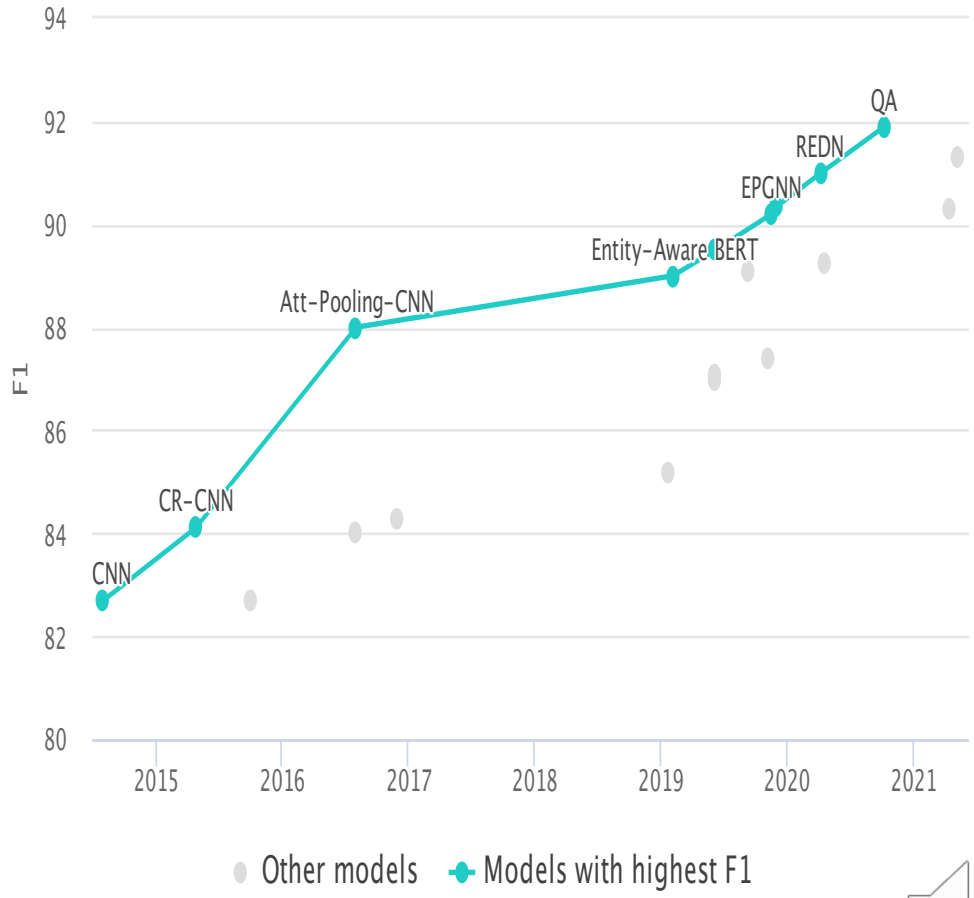
Performance on Different Datasets

Performance of Deep learning Models

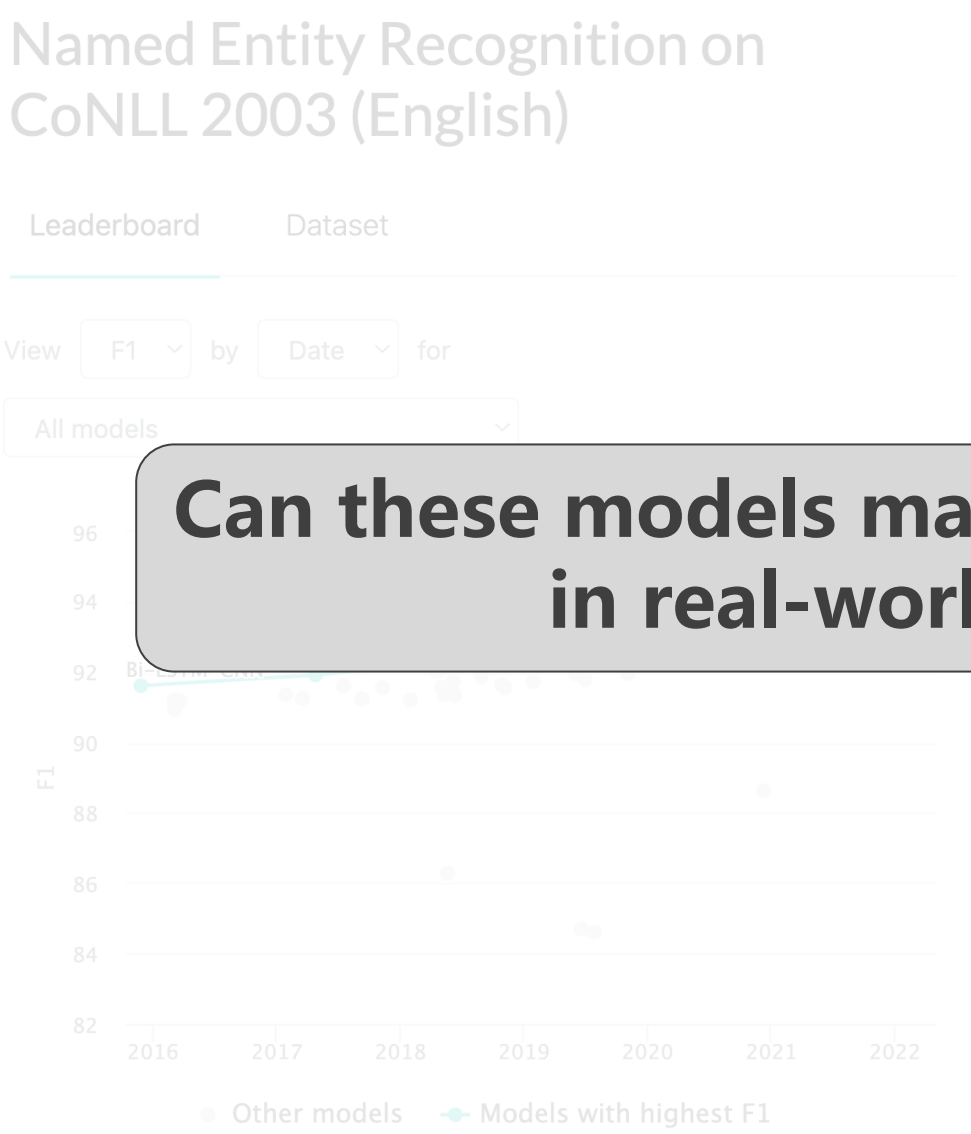
Named Entity Recognition on CoNLL 2003 (English)

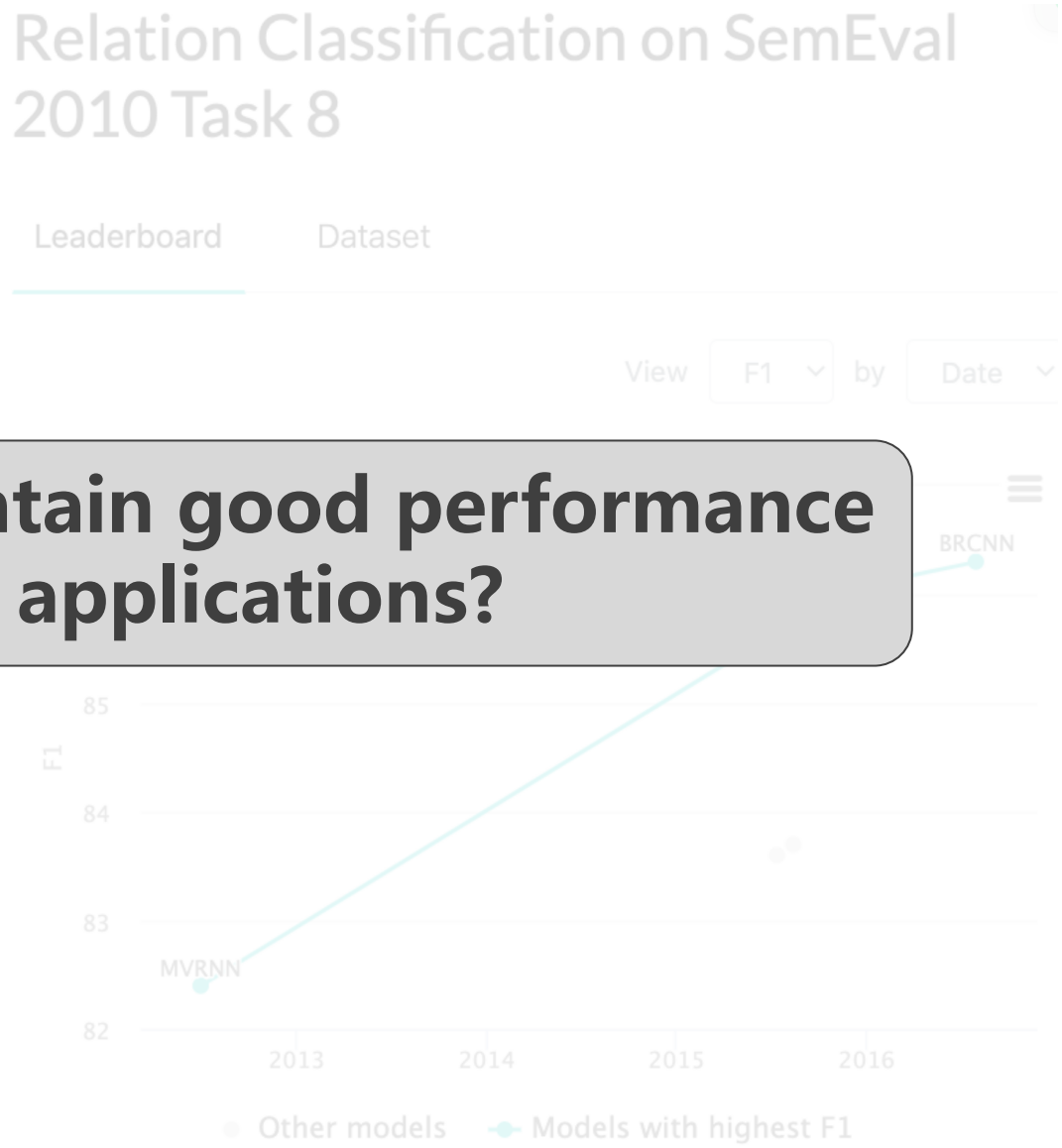


Relation Extraction on SemEval-2010 Task 8



Robustness of Information Extraction Tasks





Can these models maintain good performance in real-world applications?

Robustness of Information Extraction Tasks

EMNLP 2020

Benchmarks are blessed with strong name regularity, high mention coverage and sufficient context diversity.

When scaling NER to open situations, these advantages may no longer exist

	Regular NER	Open NER
Typical Categories	Person, Location, Organization, etc.	Movie, Song, Book, TV Series, etc.
Name Regularity	Entity types with strong regularity	Entity types with weak or no regularity
Mention Coverage	Training set with high mention coverage	Many new and unseen mentions
Context Pattern	With decent training instances to capture	Fully-annotated training data is rare
Examples	**Location** Train *starting from [**Cherry Street**]* *... at [**8th Avenue**] ...* Test ⇩ *... at [**Cherry Street**]...* *... go to [**9th Avenue**] ...*	**Movie** Train *I watched [**avatar**]last night* *...[**the matrix**] is the best...* Test ⇩ *Wow...[**Joker**] was great!* *Love [**inception**] so much.*

Figure 1: Comparison between regular NER benchmarks and open NER tasks in reality.

Lin et al., *A Rigorous Study on Named Entity Recognition: Can Fine-tuning Pretrained Model Lead to the Promised Land?*, EMNLP 2020

Robustness of Information Extraction Tasks

Settings	Name	Mention	Context	Examples
Vanilla Baseline	√	√	√	Train { *[Putin] concluded his two days of talks.* *[Blair] spoke to [Bush] on April 5.* Test *[Putin] will face re-election in March 2004.*
Name Permutation (NP)	×	√	√	Train { *[the united] concluded his two days of talks.* *[Hillsborough] spoke to [analysts] on April 5.* Test *[the united] will face re-election in March 2004.*
Mention Permutation (MP)	×	×	√	Train { *[the united] concluded his two days of talks.* *[Hillsborough] spoke to [analysts] on April 5.* Test *[which girl] will face re-election in March 2004.*
Context Reduction (CR)	√	√	↓	Train { *[Putin] concluded his two days of talks.* *[Blair] concluded his two days of talks.* *[Bush] concluded his two days of talks.* Test *[Putin] will face re-election in March 2004.*
Mention Reduction (MR)	↓	↓	√	Train { *[Blair] concluded his two days of talks.* *[Blair] spoke to [Blair] on April 5.* Test *[Putin] will face re-election in March 2004.*

Table 1: Illustration of our four kinds of randomization test. The utterances in square brackets are entity mentions. Name: name regularity knowledge; Mention: high mention coverage; Context: sufficient training instances for context diversity √: the knowledge is preserved in this setting; ×: the knowledge is erased from the data in the setting; ↓: the knowledge decreases.

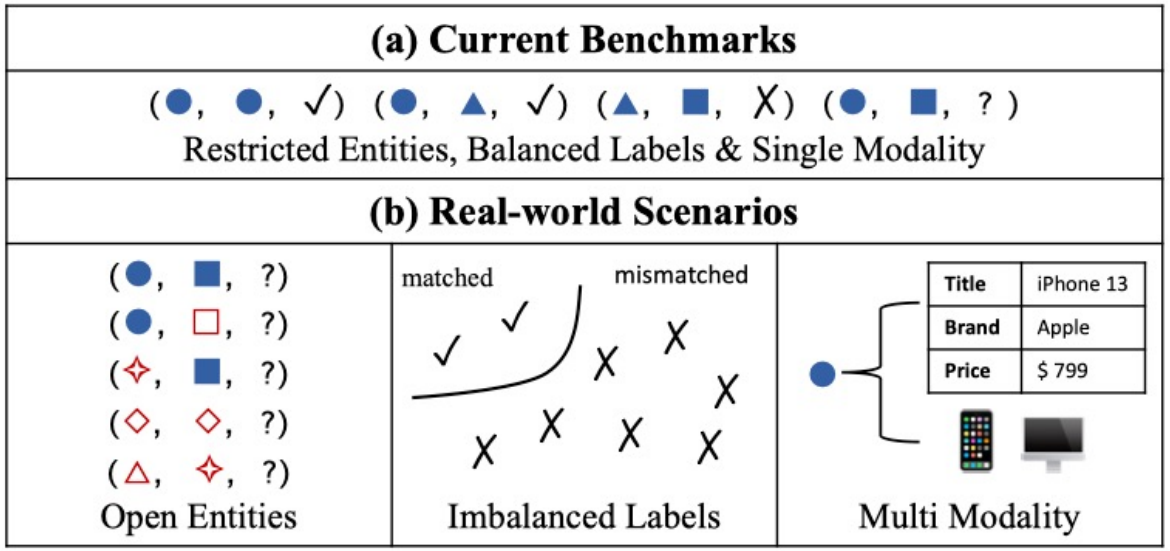
Robustness of Information Extraction Tasks

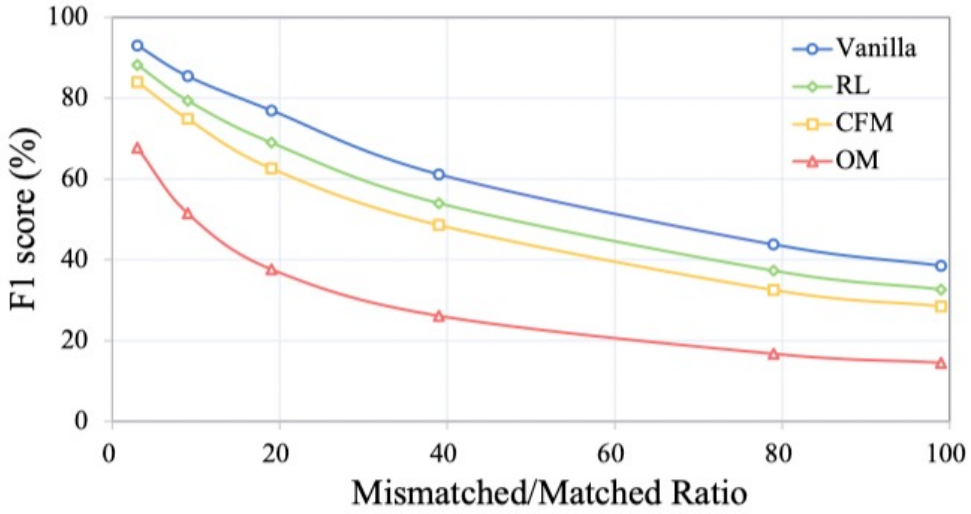
Data Setting	PER	ORG	GPE	FAC	LOC	WEA	VEH	ALL
Baseline	86.31	76.49	80.89	69.23	40.58	74.70	61.97	81.76
Name Permutation	73.41	44.34	49.71	37.96	28.24	33.33	23.93	62.28
- Drop Compared with Baseline	15%	42%	39%	45%	44%	55%	61%	24%
Mention Permutation	61.78	39.40	33.27	32.16	18.60	9.38	21.92	51.58
- Drop Compared with Baseline	28%	48%	59%	54%	54%	87%	65%	34%

Table 2: Micro-F1 scores of BERT-CRF tagger on original data, name permutation setting and mention permutation setting respectively. We can see that erasing name regularity and mention coverage will significantly undermine the model performance.

Lin et al., *A Rigorous Study on Named Entity Recognition: Can Fine-tuning Pretrained Model Lead to the Promised Land?*, EMNLP 2020

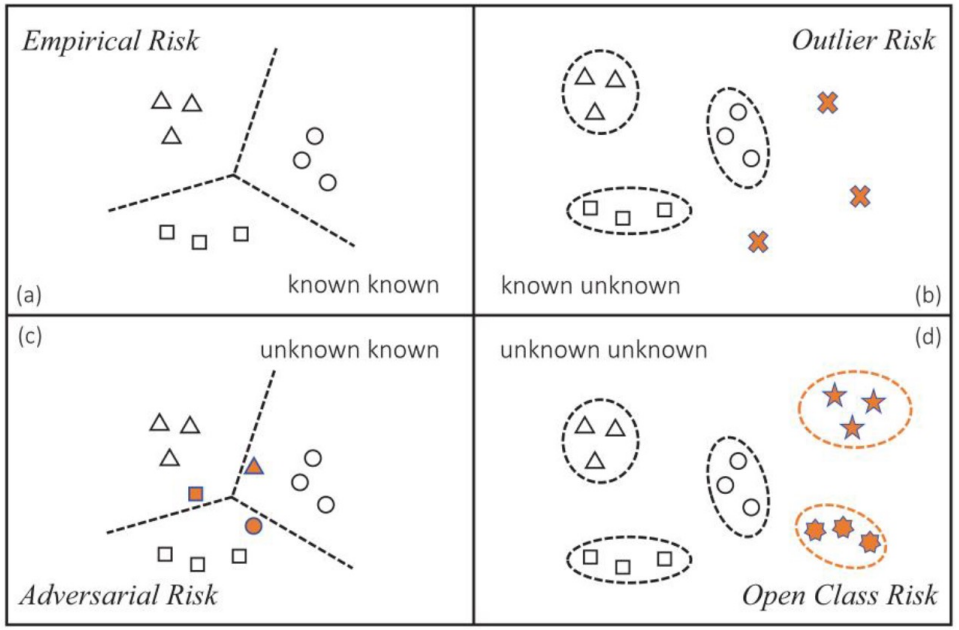
Robustness of Information Extraction Tasks





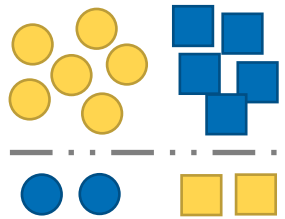
wang et al., Bridging the Gap between Reality and Ideality of Entity Matching, IJCAI 2022

Reasons for Robustness Problem

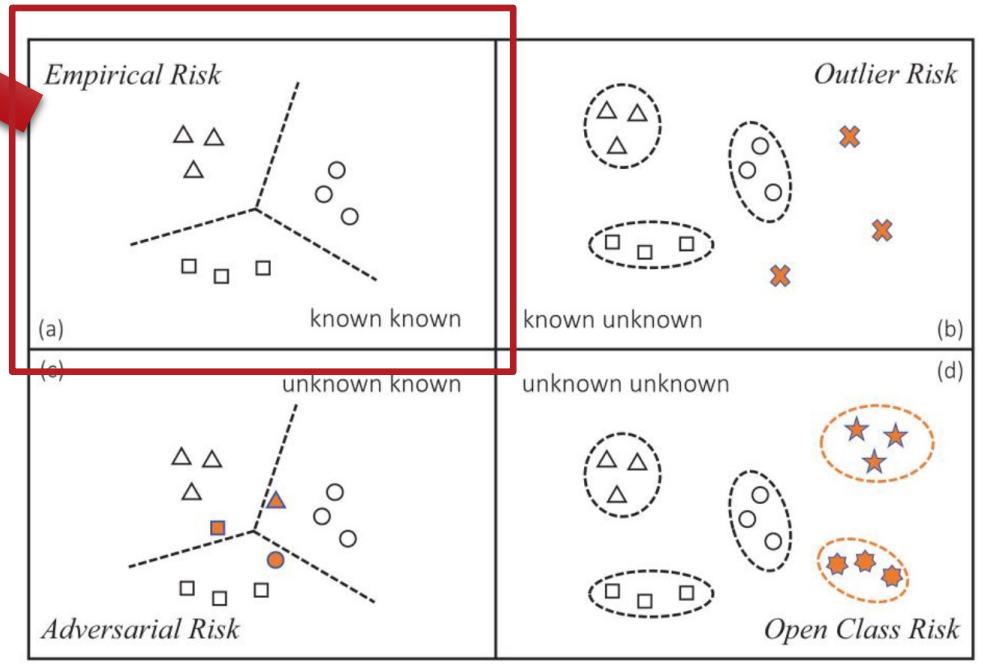


Zhang X Y, Liu C L, Suen C Y. Towards robust pattern recognition: A review[J]. Proceedings of the IEEE, 2020, 108(6): 894-922.

Bias Problems in Knowledge Graph Construction

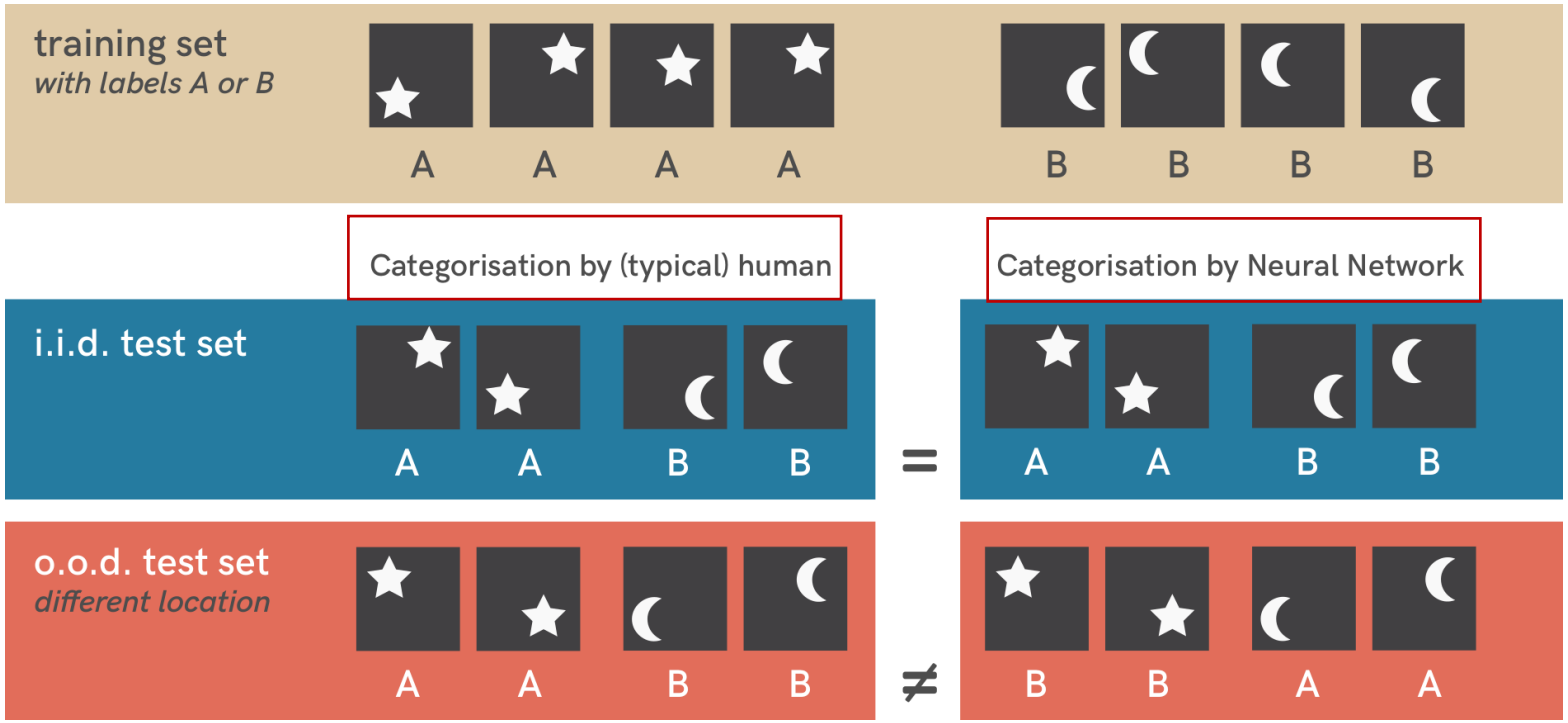


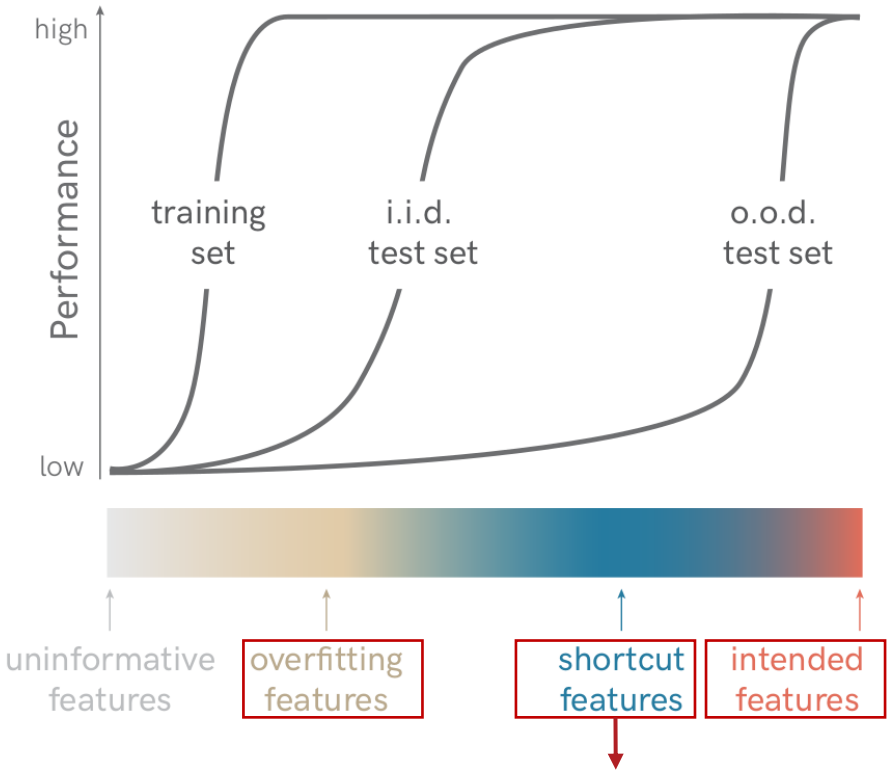
Bias Data



Zhang X Y, Liu C L, Suen C Y. Towards robust pattern recognition: A review[J]. Proceedings of the IEEE, 2020, 108(6): 894-922.

Bias Problems in Knowledge Graph Construction





Star – top right corner or bottom left corner

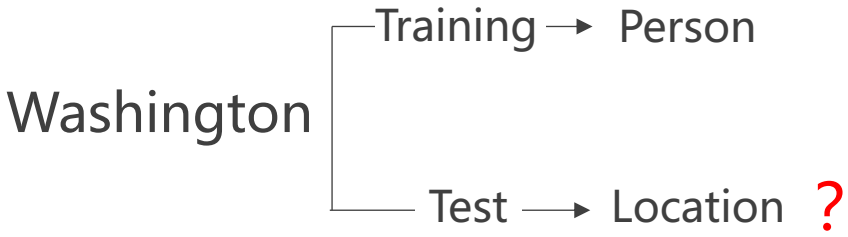
Moon – top left corner or bottom right corner

Problems of Robustness

Geirhos, Robert, et al. "Shortcut learning in deep neural networks." *Nature Machine Intelligence* 2.11 (2020): 665-673

Bias Problems in Knowledge Graph Construction

Datasets	Embed-layer		Entity Coverage Rate					
	Char	Word	Overall	1	$(0.5, 1)$	$(0, 0.5]$	$C \neq 0$	$C = 0$
CoNLL	CNN	-	76.42	79.94	86.99	78.84	69.74	77.61
	FLAIR	-	89.98	95.30	95.58	82.39	72.16	90.39
	ELMo	-	91.79	97.61	95.98	85.15	71.43	92.22
	BERT	-	91.34	97.72	95.17	86.66	**77.83**	92.37
	-	Rand	78.43	95.05	94.75	73.54	37.97	66.40
	-	GloVe	89.10	98.44	96.31	81.34	57.80	87.23
	CNN	Rand	82.88	94.13	94.48	74.25	47.78	78.91
	CNN	GloVe	90.33	98.32	95.94	80.33	59.67	89.74
	ELMo	GloVe	92.46	98.08	**96.46**	86.14	69.79	93.08
	FLAIR	GloVe	**93.03**	**98.56**	96.38	**87.07**	73.58	**93.42**
WNUT	CNN	-	20.88	45.99	67.01	40.25	19.14	19.74
	FLAIR	-	41.49	81.15	88.14	54.36	39.56	43.44
	ELMo	-	43.70	88.72	90.83	55.56	**44.19**	43.32
	BERT	-	44.08	77.75	81.61	49.74	34.65	41.92
	-	Rand	14.97	60.62	83.84	50.00	3.90	4.77
	-	GloVe	37.28	89.29	92.62	45.65	35.34	35.15
	CNN	Rand	22.29	48.88	71.43	39.08	16.75	18.83
	CNN	GloVe	40.72	86.12	**92.24**	49.74	26.67	40.06
	ELMo	GloVe	45.33	90.38	89.92	56.57	37.8	46.58
	FLAIR	GloVe	**45.96**	**90.52**	89.92	**61.69**	42.07	**48.38**



Entity Coverage Ratio (ECR) The measure entity coverage ratio is used to describe the degree to which entities in the test set have been seen in the training set with the same category.

$$\rho(e_i) = \begin{cases} 0 & C = 0 \\ (\sum_{k=1}^{K} \frac{\#(e_i^{tr,k})}{C^{tr}} \#(e_i^{te,k}))/C^{te} & \text{otherwise} \end{cases} \quad (1)$$

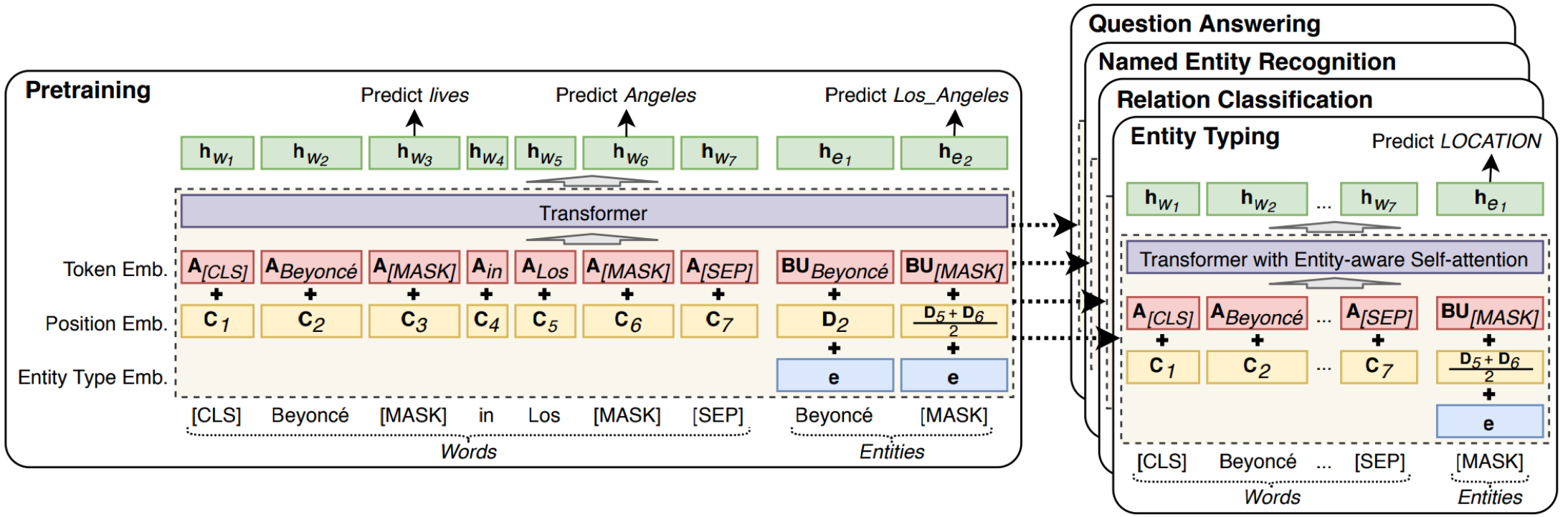
where $e_i^{tr,k}$ is the entity e_i in the training set with ground truth label k, $e_i^{te,k}$ is the entity e_i in the test set with ground truth label k, $C^{tr} = \sum_{k=1}^{K} \#(e_i^{tr,k})$, $C^{te} = \sum_{k=1}^{K} \#(e_i^{te,k})$, and $\#$ denotes the counting operation.

Fu et al. , *Rethinking Generalization of Neural Models: A Named Entity Recognition Case Study* , *AAAI 2020*

Bias Problems in Knowledge Graph Construction

How to alleviate the bias Problem ?

1. Additional Knowledge ： Introduce entity knowledge in the pre-training



Yamada, Ikuya, et al. “Luke: deep contextualized entity representations with entity-aware self-attention.” EMNLP 2020.

Bias Problems in Knowledge Graph Construction

How to alleviate the bias Problem ?

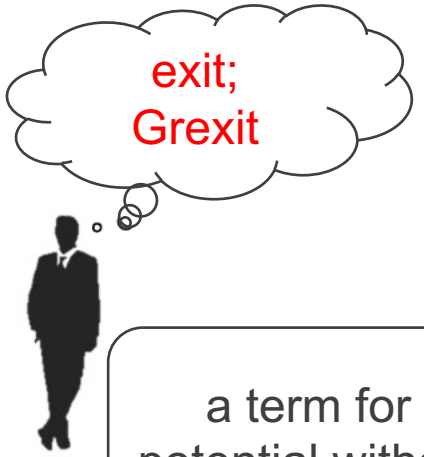
2. Semantic Reasoning： Estimate OOV word representations using existing similar words

Why **Brexit** needs renegotiating?

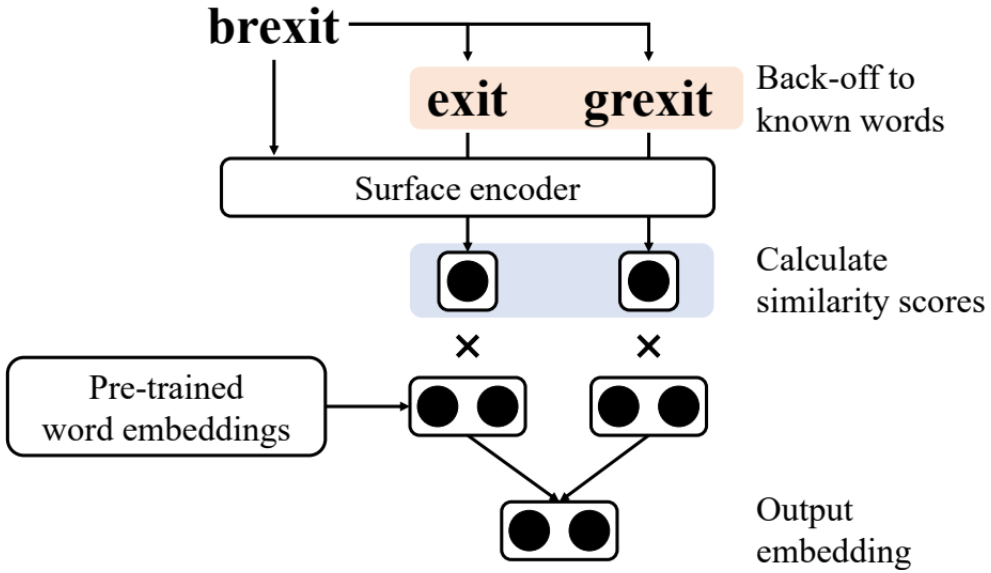
What is **Brexit** ?

event ?

organization ?



a term for the potential withdrawal

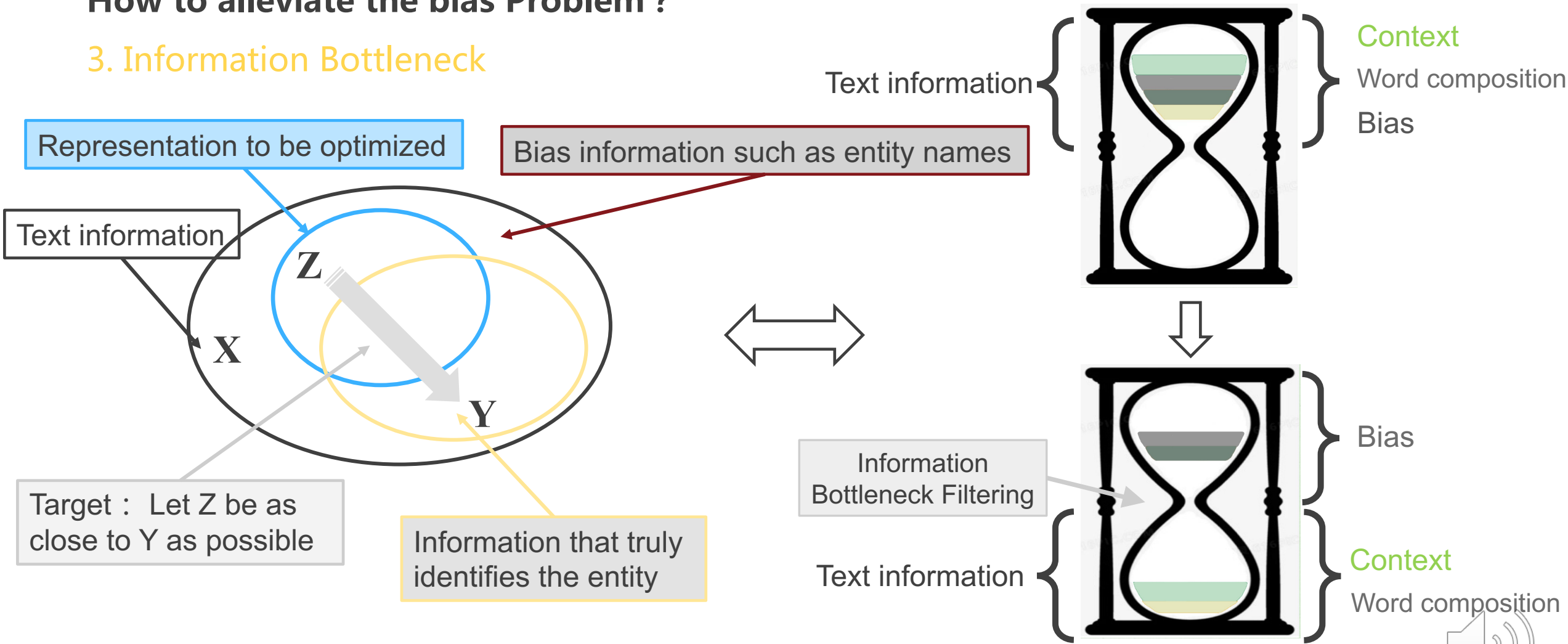


Fukuda, Nobukazu, Naoki Yoshinaga, and Masaru Kitsuregawa. "Robust Backed-off Estimation of Out-of-Vocabulary Embeddings." EMNLP2020.

Bias Problems in Knowledge Graph Construction

How to alleviate the bias Problem ?

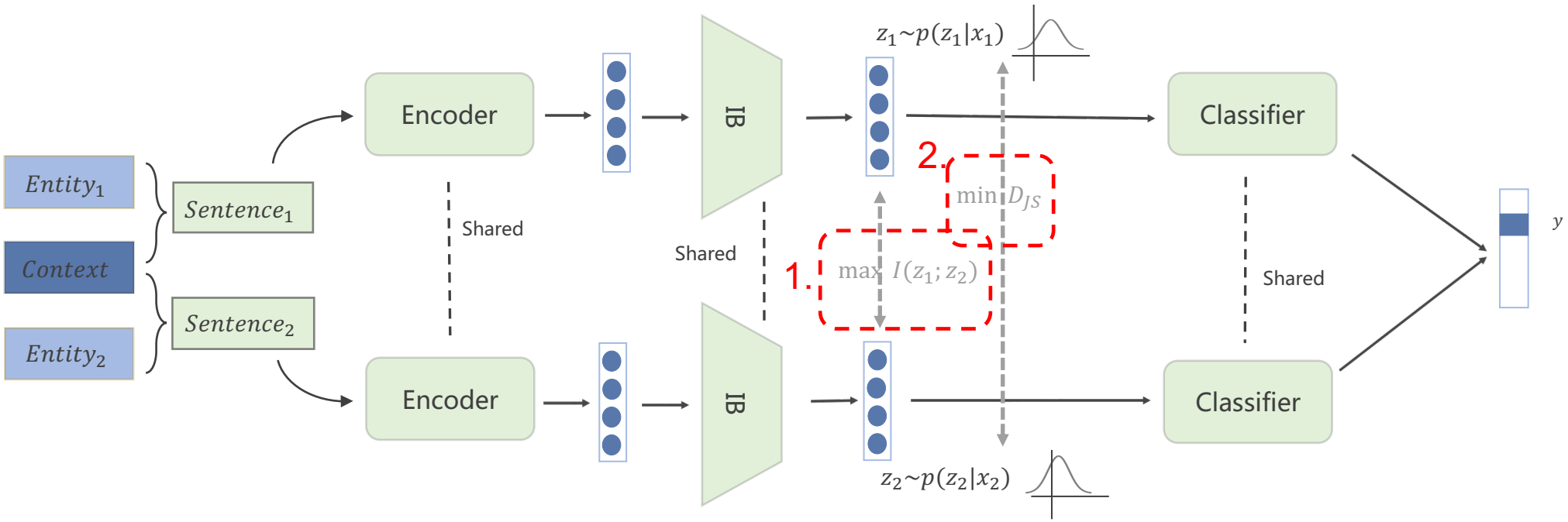
3. Information Bottleneck



Wang et al., MINER: Improving Out-of-Vocabulary Named Entity Recognition from an Information Theoretic Perspective, ACL 2022

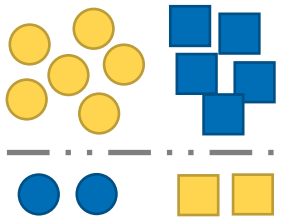
Bias Problems in Knowledge Graph Construction

- Suppose x_1 and x_2 are two samples with the same context, but with different entities.
  - Shanghai Oriental Pearl Tower is located in Lujiazui, Shanghai.
  - Jin Mao Tower is located in Lujiazui, Shanghai.

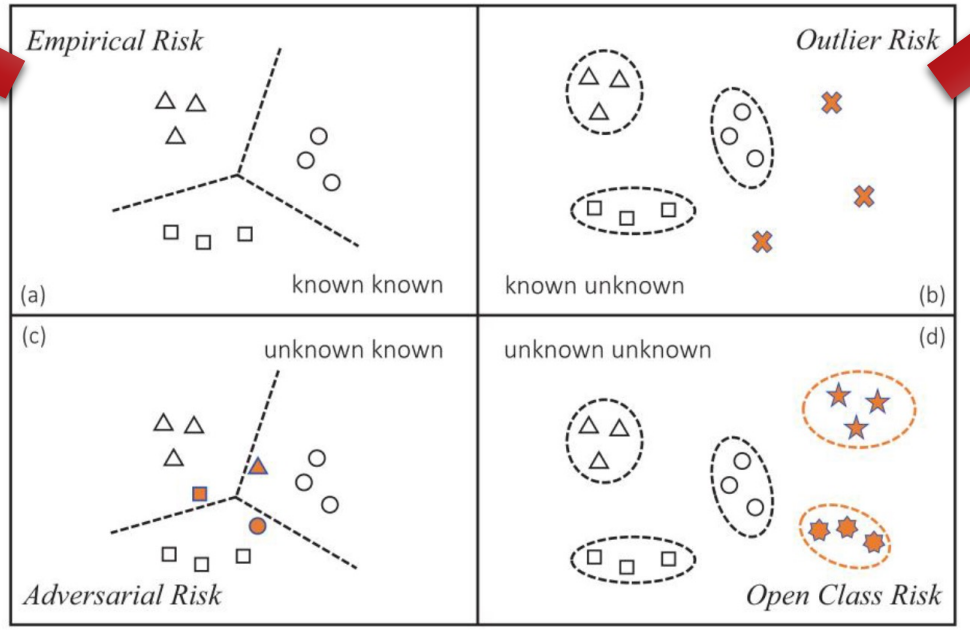


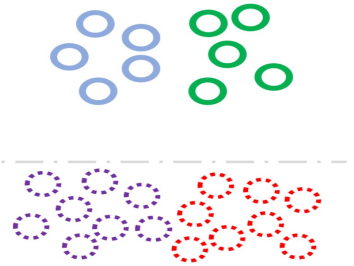
Wang et al., MINER: Improving Out-of-Vocabulary Named Entity Recognition from an Information Theoretic Perspective, ACL 2022

Noisy label Problems in Knowledge Graph Construction



Bias Data

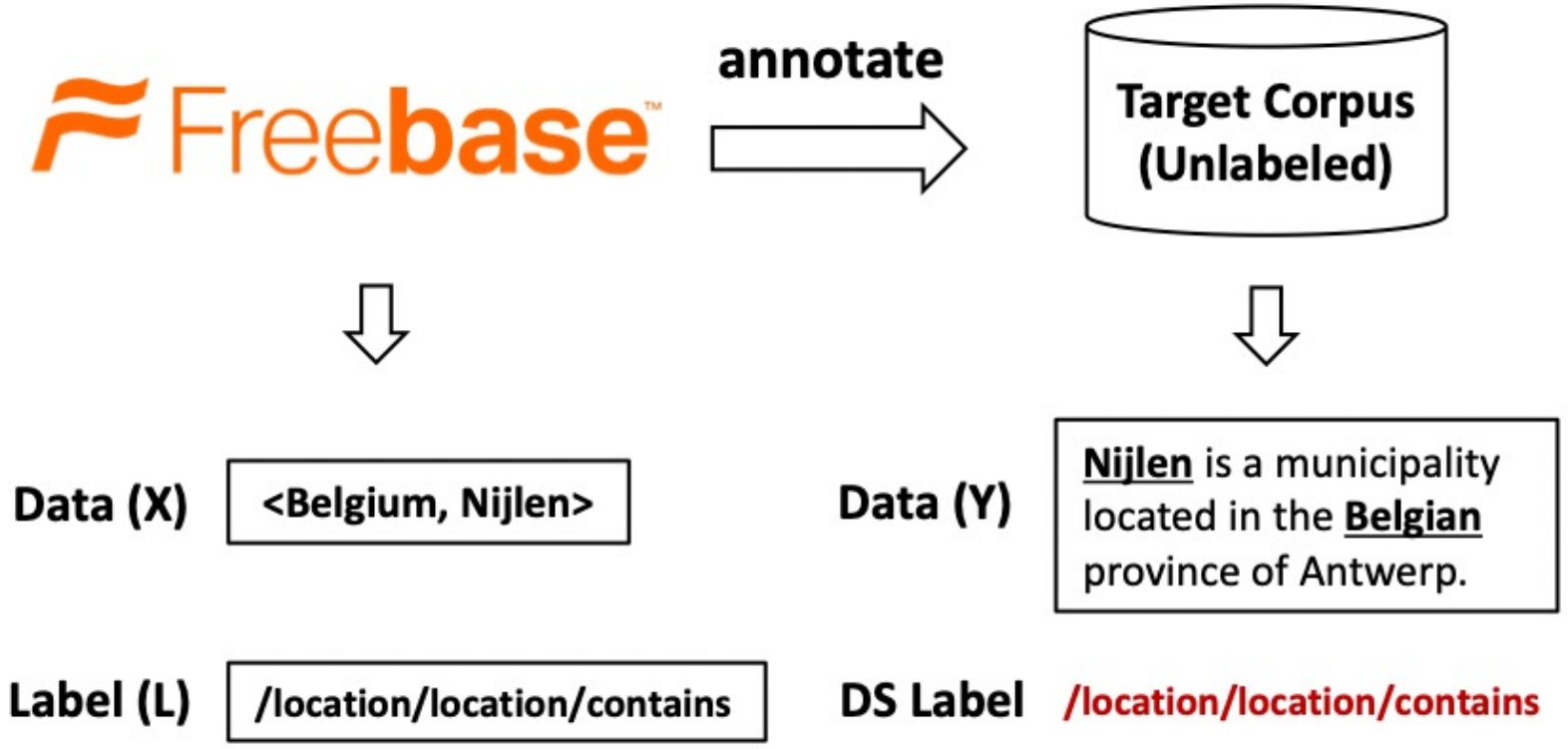




Noisy Data

Noisy label Problems in Knowledge Graph Construction

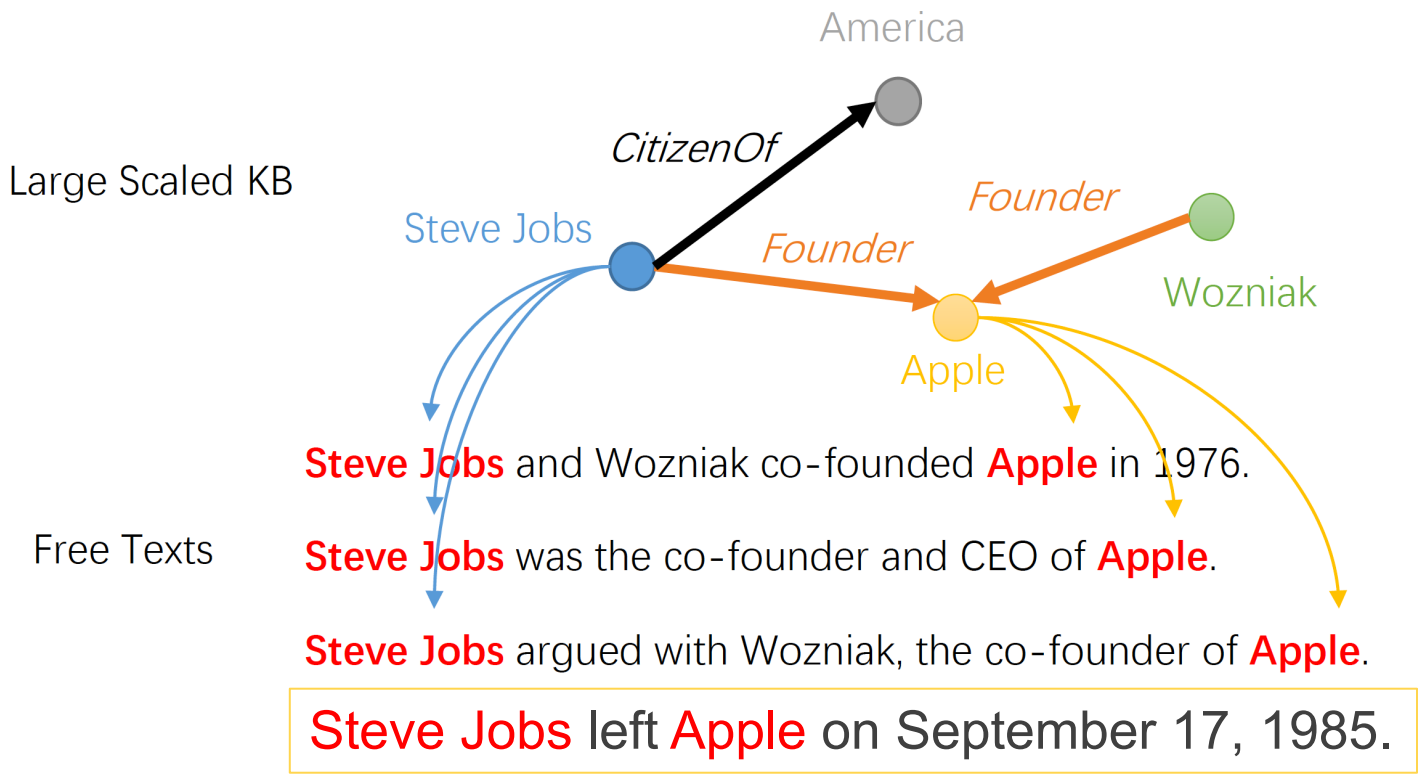
distant supervisd Relation Extraction



Jiang J, Allan J. Similarity-based distant supervision for definition retrieval. CIKM. 2017: 527-536.

Noisy label Problems in Knowledge Graph Construction

- **There is a lot of noise in distant supervision**



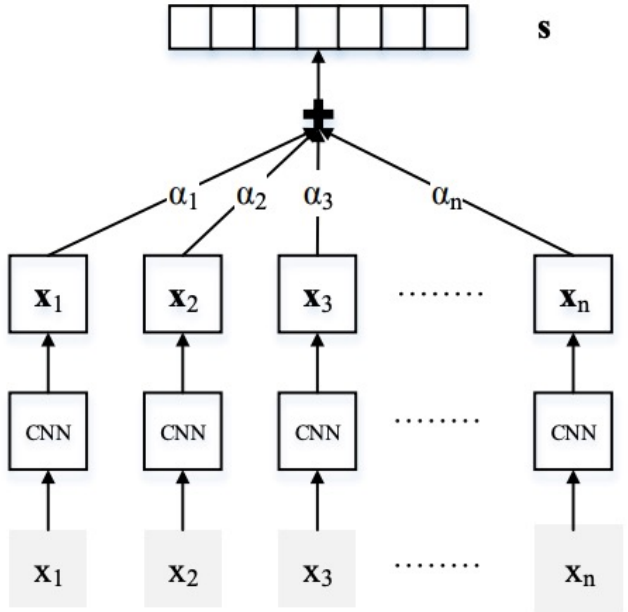
Noisy label Problems in Knowledge Graph Construction

➢ **How to alleviate the noise Problem** ?

- Suppose that at least one sentence expresses the relation-> Multi-Instance Learning（Bag-level）

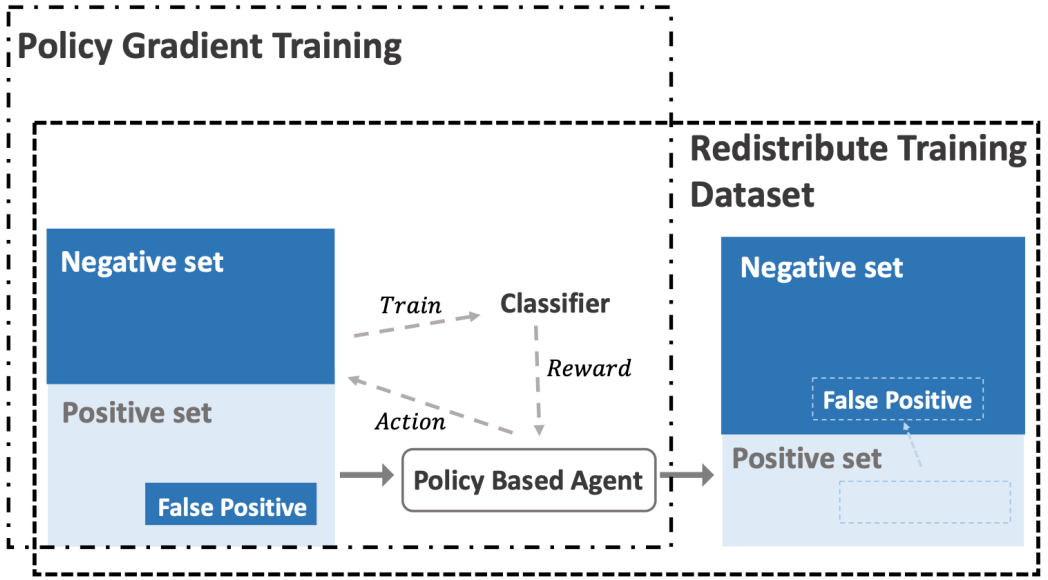
➢ Attention-based

- PCNN+ATT



➢ Reinforcement learning

- RL-based noise selection



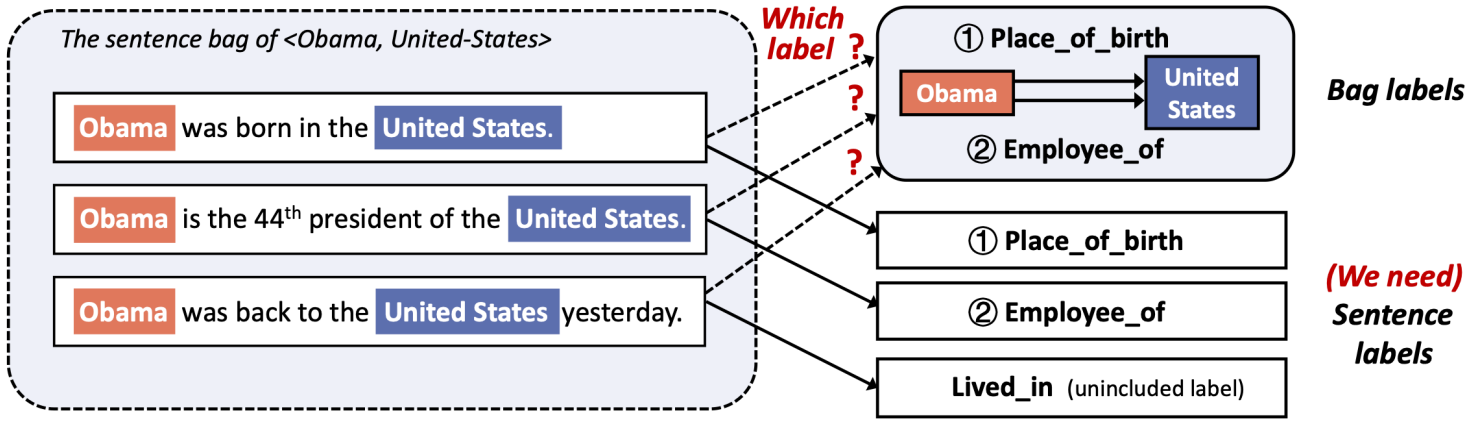
"Neural relation extraction with selective attention over instances." ACL2016.

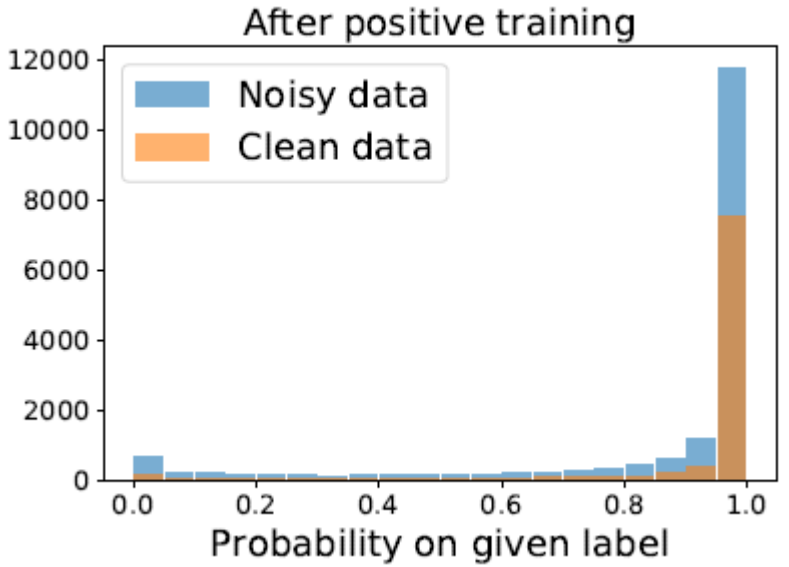
"Robust distant supervision relation extraction via deep reinforcement learning."ACL2018

Noisy label Problems in Knowledge Graph Construction

Important for downstream tasks

- Sentence-level label cannot be obtained using bag-level relation extraction
- There are many types of noise
- Traditional positive learning methods cannot effectively filter noisy data





Noisy label Problems in Knowledge Graph Construction

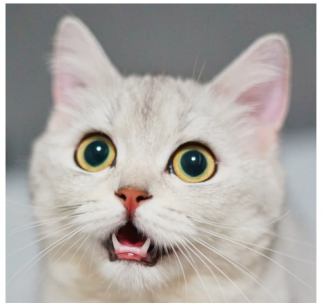
Positive Training

"is a **dog**" — Correct Labels

"is a **cat**"
"is a **bird**"
"is an **elephant**"
"is a **chicken**"
"is a **pig**" — Incorrect labels

$$\mathcal{L}_{PT}(f, y^*) = -\sum_{k=1}^{C} y_k \log p_k$$

Negative Training



Correct Labels:
"is not a **dog**"
"is not a **bird**"
"is not an **elephant**"
"is not a **chicken**"
"is not a **pig**"

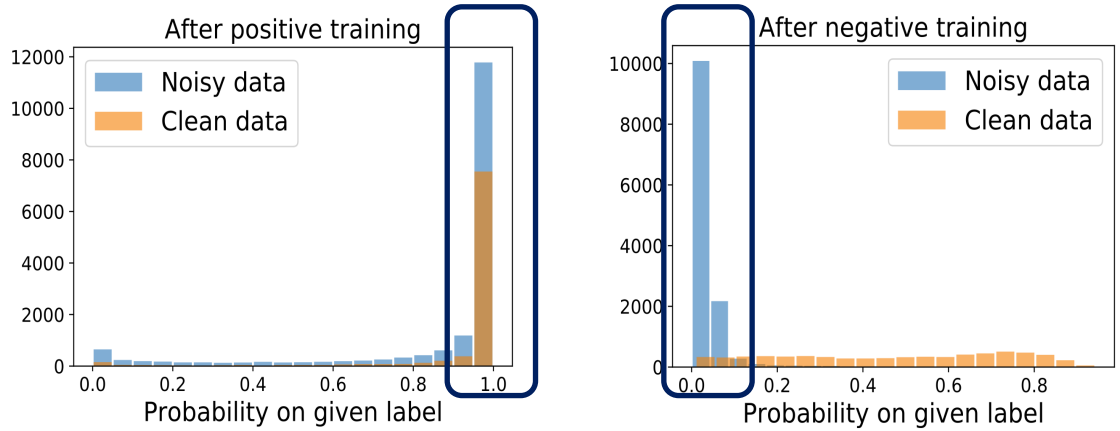
Incorrect labels:
"is not a **cat**"

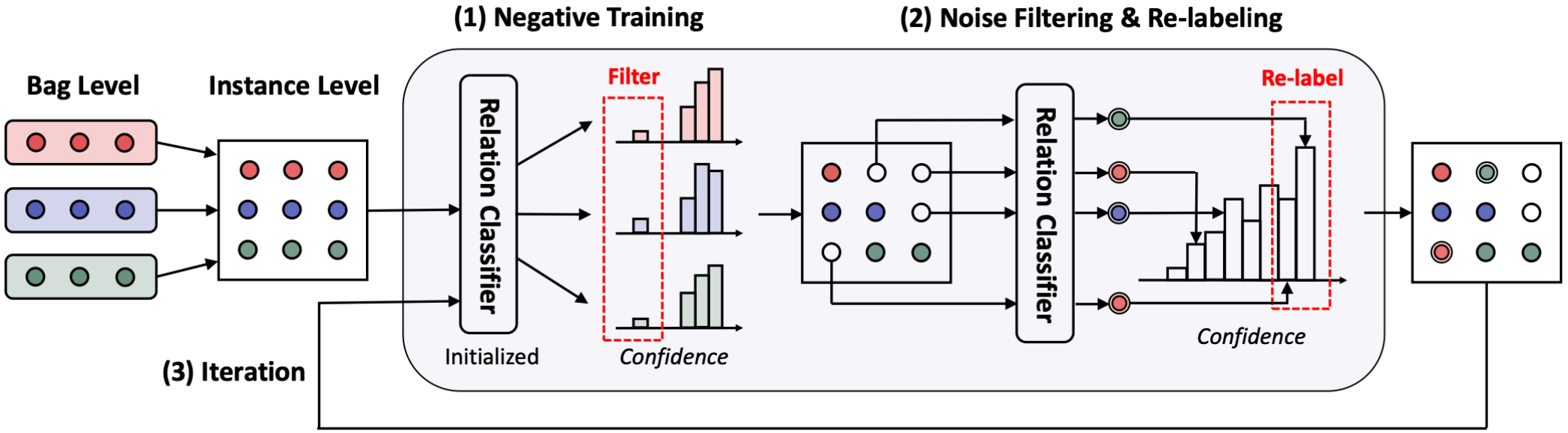
$$\mathcal{L}_{NT}(f, y^*) = -\sum_{k=1}^{C} \overline{y_k} \log(1 - p_k)$$

Noisy label Problems in Knowledge Graph Construction

Our methods :

- Negative learning for distinguishing noisy data
- Data relabeling transforms noisy data into useful training data
- Iterative training for further enhancing performance





SENT: Sentence-level Distant Relation Extraction via Negative Training. ACL 2021

Noisy label Problems in Knowledge Graph Construction

323 Samples including 200 incorrect samples

Noise Reduction	Prec.	Rec.	F1
CNN+RL_2	40.58	96.31	57.10
ARNOR	76.37	68.13	72.02
SENT (biLSTM)	80.00	**88.46**	84.02
SENT (biLSTM+BERT)	**84.33**	85.67	**84.99**

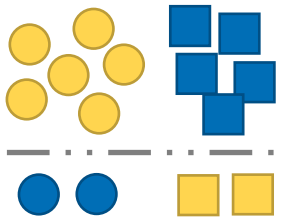
Table 3: The noise-filtering effect evaluated on a noise-annotated test set of NYT-10.

13012 correct samples
20586 incorrect samples

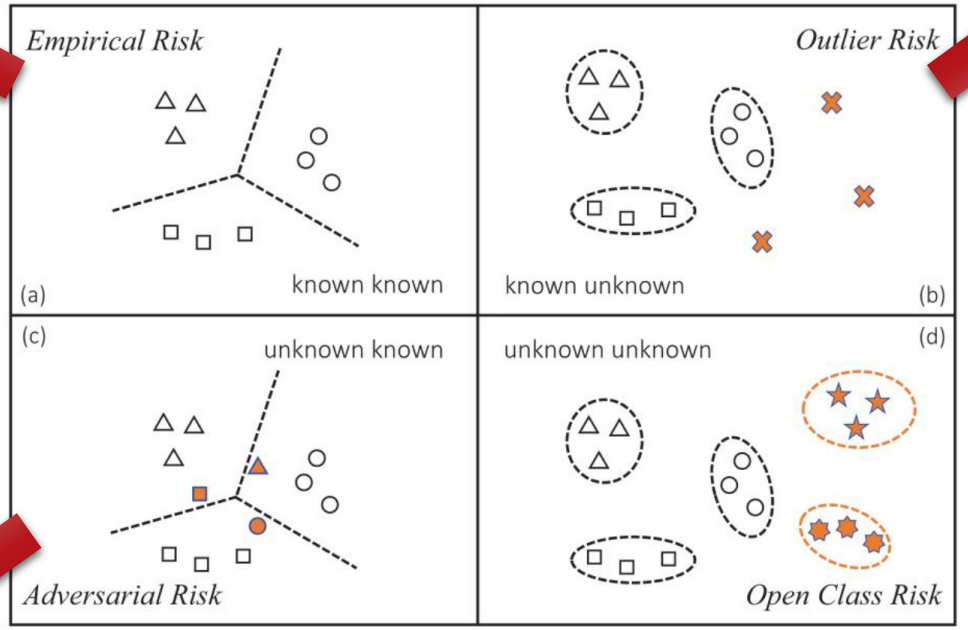
	Method	Prec.	Rec.	F1
Clean Data	BiLSTM+ATT	67.7	63.2	65.4
	BiLSTM	61.4	61.7	61.5
Noisy Data	BiLSTM+ATT	32.8	43.8	37.5
	BiLSTM	37.8	45.5	41.3
	SENT (biLSTM)	**66.0**	**52.9**	**58.7**

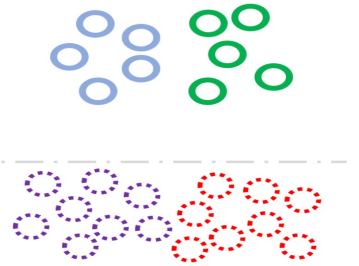
Table 4: Model performance on clean and noisy-TACRED. When trained on noisy data, the performance of base models degrade dramatically while SENT achieves comparable results with the models trained on clean data.

Adversarial Sample Problems in Knowledge Graph Construction

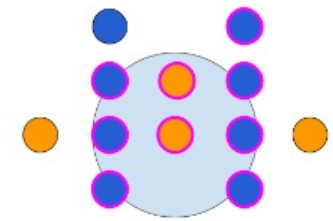


Bias Data



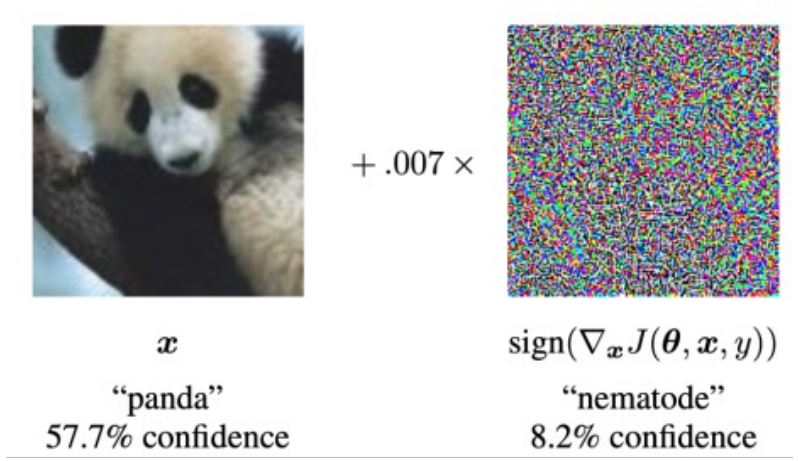


Noisy Data

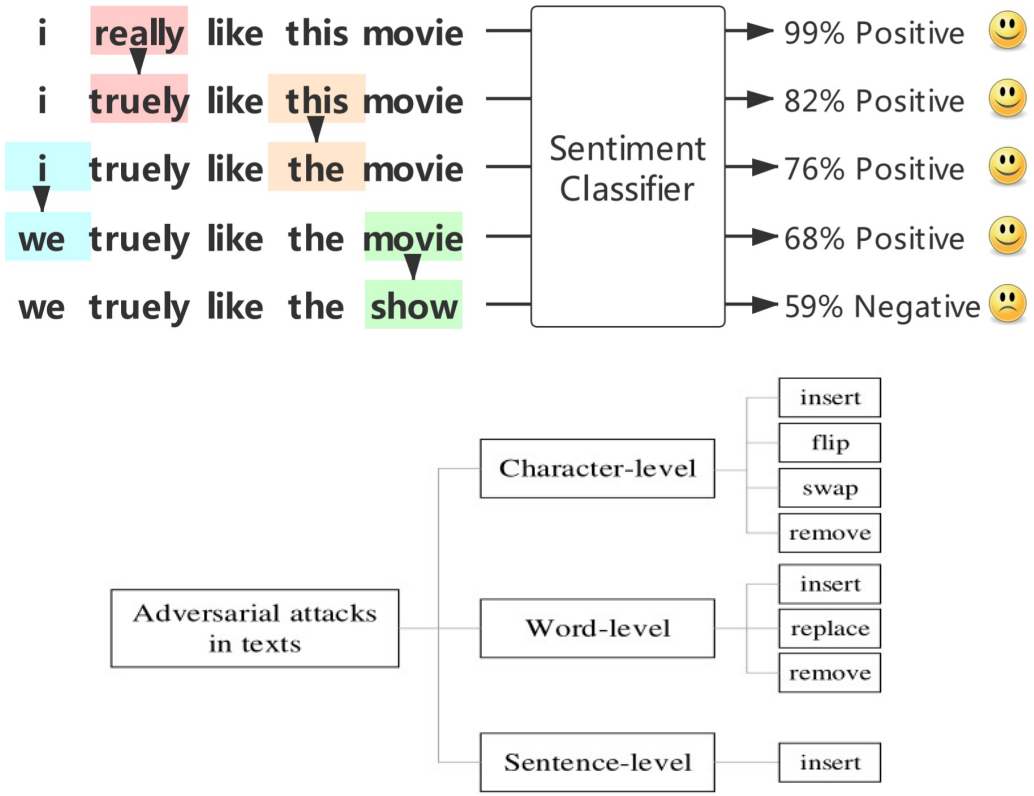


Adversarial example

Adversarial Attack



$$\mathbf{x}' = \mathbf{x} + \eta, f(\mathbf{x}) = \mathbf{y}, \mathbf{x} \in \mathbf{X}$$
$$f(\mathbf{x}') \neq \mathbf{y}$$
$$\text{or } f(\mathbf{x}') = \mathbf{y}', \mathbf{y}' \neq \mathbf{y}$$

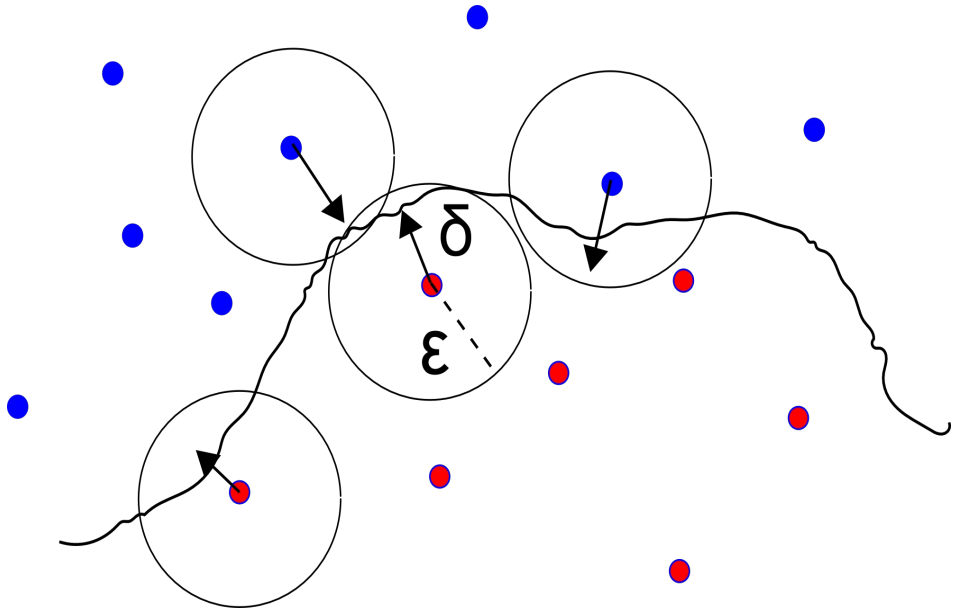


C. Szegedy et al., Intriguing properties of neural networks, ICLR 2014.

Adversarial Sample Problems in Knowledge Graph Construction

Adversarial Training：

- Adding adversarial samples to model training can improve the generalization ability and robustness of the model.



Adversarial Sample Problems in Knowledge Graph Construction

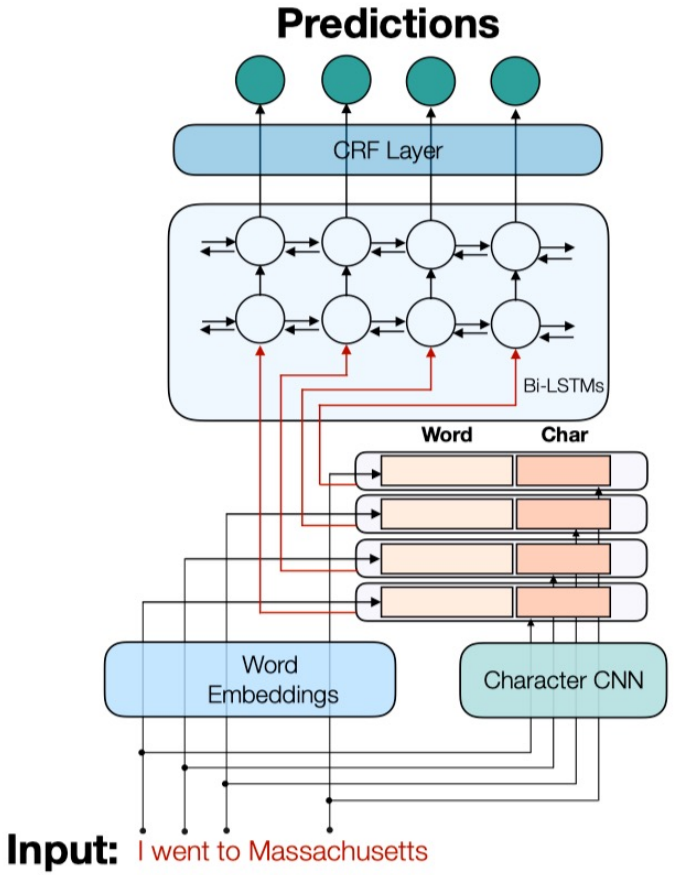


Figure 1: Sequence Labeling Model Architecture.

Adversarial Training

$$d_w = \underset{\epsilon, ||\epsilon||_2 \leq \delta_w}{\text{argmax}} \; Loss(y; w + \epsilon, c, \hat{\theta})$$

Virtual Adversarial Training

$$\underset{\epsilon, ||\epsilon||_2 \leq \delta_w}{\text{argmax}} \; KL(P(\hat{y}; w, c, \hat{\theta}) || P(\hat{y}; w + \epsilon, c, \hat{\theta}))$$

Sequence Virtual Adversarial Training

$$P'(S; w, c, \hat{\theta}) = (p'_1, p'_2, .., p'_k, 1 - \sum_{i=1} p'_i) \quad (15)$$

$$where \quad p'_i = p_{crf}(s_i; w, c, \hat{\theta}), i \in [1, k]$$

$$P_{adv} = P'(S; w + d_w, c + d_c, \hat{\theta})$$

Chen et al. SeqVAT: Virtual adversarial training for semi-supervised sequence labeling, ACL 2020.

Adversarial Sample Problems in Knowledge Graph Construction

Challenges of Adversarial Training in NLP:

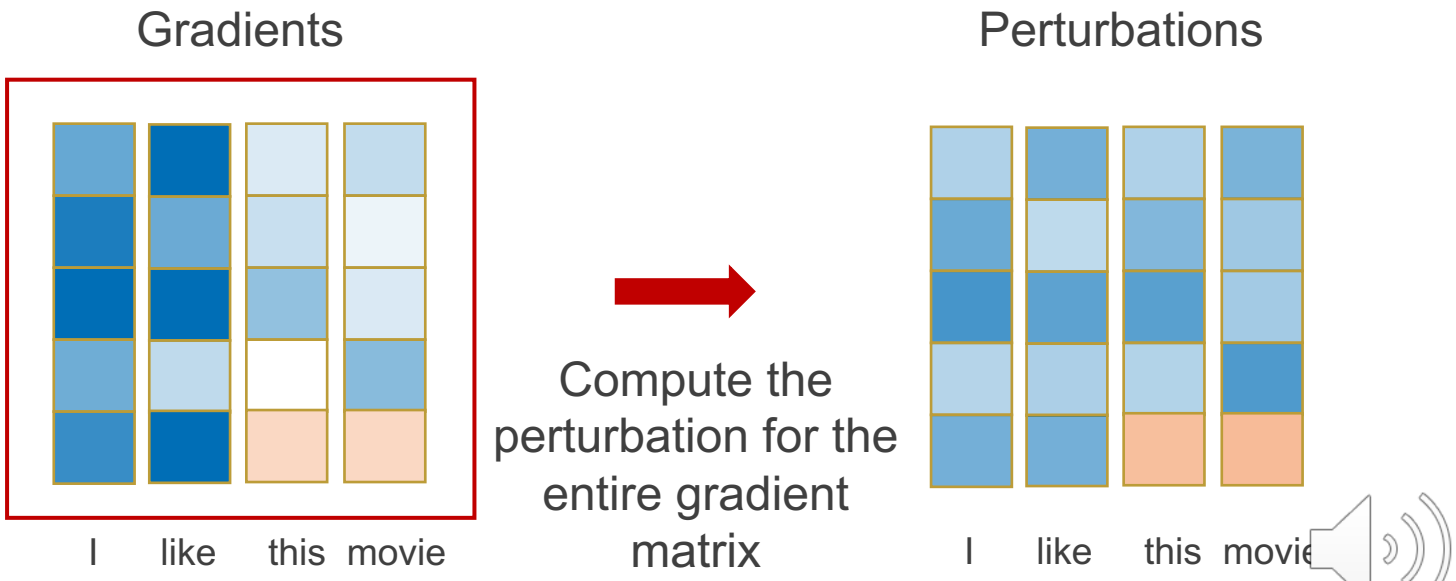
- Discreteness of natural language sequences

Problems with traditional adversarial training:

- Randomly initializing the perturbations on the same tokens in different sequences
- Instance-level perturbation on the embeddings of the entire sequence.

$$\boldsymbol{\delta}_{t+1} = \prod_{||\boldsymbol{\delta}_t||_F \leq \epsilon} \frac{(\boldsymbol{\delta}_t + \alpha g(\boldsymbol{\delta}_t))}{||g(\boldsymbol{\delta}_t)||_F}$$

$$g(\boldsymbol{\delta}_t) = \nabla_{\boldsymbol{\delta}} L(f_\theta(\boldsymbol{X} + \boldsymbol{\delta}_t), y)$$



TAVAT: Token-Aware Virtual Adversarial Training for Language Understanding, AAAI 2021

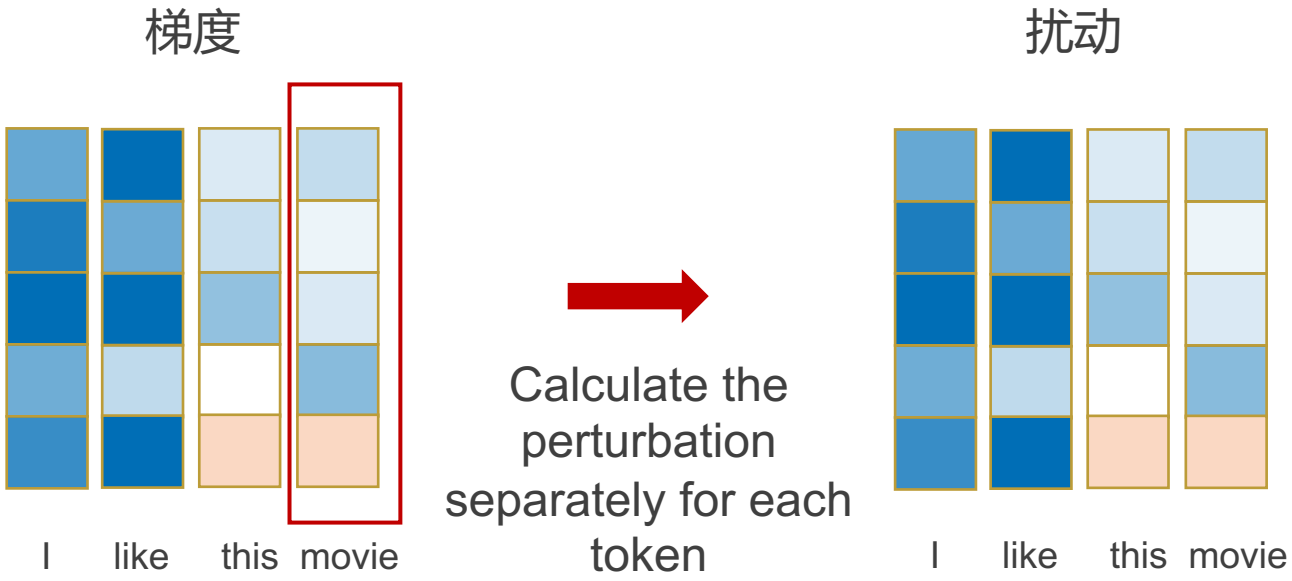
Adversarial Sample Problems in Knowledge Graph Construction

Improvement：

- Perturbation is calculated separately for each token,
- preserving the discrete nature of the sequence

$$\eta_{t+1}^i = \eta^i * \frac{(\eta_t^i + \alpha g(\eta_t^i))}{||g(\eta_t^i)||_F)}$$

$$\eta_{t+1} = \prod_{||\eta||_F \leq \epsilon}(\eta_t)$$



TAVAT: Token-Aware Virtual Adversarial Training for Language Understanding, AAAI 2021

Adversarial Sample Problems in Knowledge Graph Construction

Results :

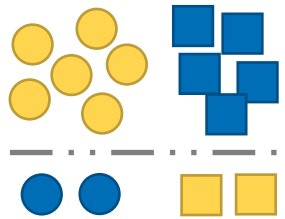
1. Better accuracy on different task
2. Better robustness on different tasks

Model	RTE	QNLI	MRPC	CoLA	SST	STS-B	MNLI-m/mm	QQP
	Acc	Acc	Acc/f1	Mcc	Acc	P/S Corr	Acc	Acc/f1
BERT-BASE								
BERT (Devlin et al. 2018)	-	88.4	-/86.7	-	92.7	-	84.4/-	-
BERT-ReImp	63.5	91.1	84.1/89.0	54.7	92.9	89.2/88.8	84.5/84.4	90.9/88.3
FreeAT-ReImp	68.0	91.3	85.0/89.2	57.5	93.2	89.5/89.0	84.9 / 85.0	91.2/88.5
FreeLB-ReImp	70.0	91.5	86.0/90.0	58.9	93.4	89.7/89.2	85.3 / 85.5	91.4/88.6
TA-VAT(ours)	**74.0**	**92.4**	**88.0/91.6**	**62.0**	**93.7**	**90.0/89.6**	**85.7 / 85.8**	**91.6/88.9**
ALBERT-xxlarge-v2								
ALBERT-xxlarge-v2(Lan et al. 2019)	89.2	95.3	-/90.9	71.4	96.9(96.5)	93.0/-	90.8/-	92.2/-
FreeLB(Zhu et al. 2020)	89.9	95.6*	-/92.4	73.1	97.0	93.2/-	90.9/-	92.5/-
TA-VAT(ours)	**90.3**	**95.7**	-/ **93.4**	**74.1**	96.8	**93.4**/-	**91.1**/-	**92.6** /-

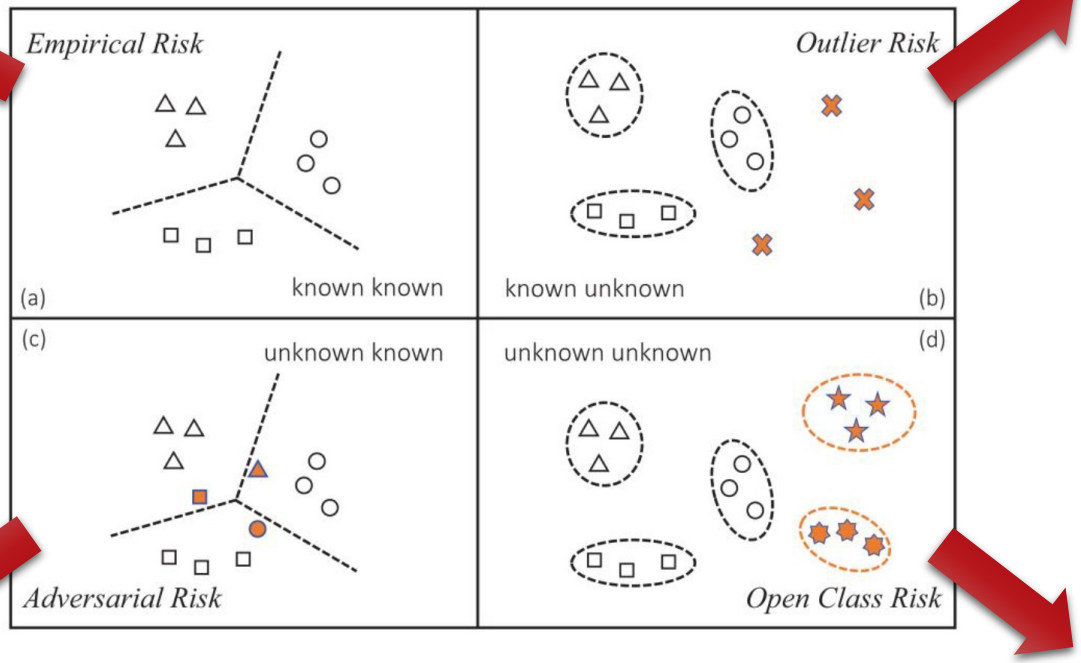
Table 1: Evaluation results on the development set of GLUE benchmark. QNLI* in FreeLB is formed as pairwise ranking task.

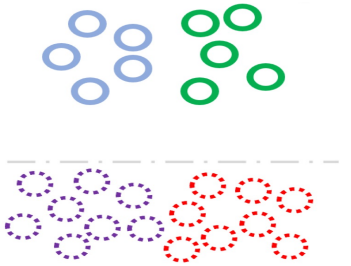
Model	RTE	QNLI	MRPC	CoLA	SST	STS-B	MNLI-m/mm	QQP
	Acc	Acc	Acc/f1	Mcc	Acc	P/S Corr	Acc	Acc/f1
BERT-BASE(Devlin et al. 2018)	66.4	90.5	88.9/84.8	52.1	93.5	87.1/85.8	84.6/83.4	71.2/89.2
FreeLB(Zhu et al. 2020)	70.1	91.8*	88.1/83.5	54.5	93.6	87.7/86.7	85.7/84.6	72.7/89.6
TA-VAT(ours)	**71.0**	**91.7**	**88.9/84.5**	**55.9**	**94.5**	86.8/85.7	85.2/**84.7**	**72.8**/89.5

Adversarial Sample Problems in Knowledge Graph Construction

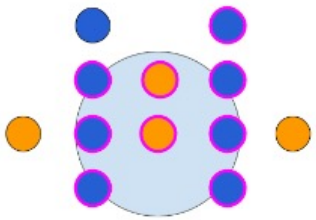


Bias Data

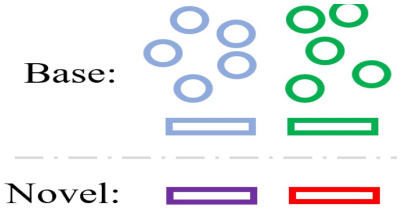




Noisy Data



Adversarial Samples



Open IE

Open IE Problems in Knowledge Graph Construction

Closed World Assumption

Training Phase

Inspections for diseases 0.15

foods for diseases 0.2

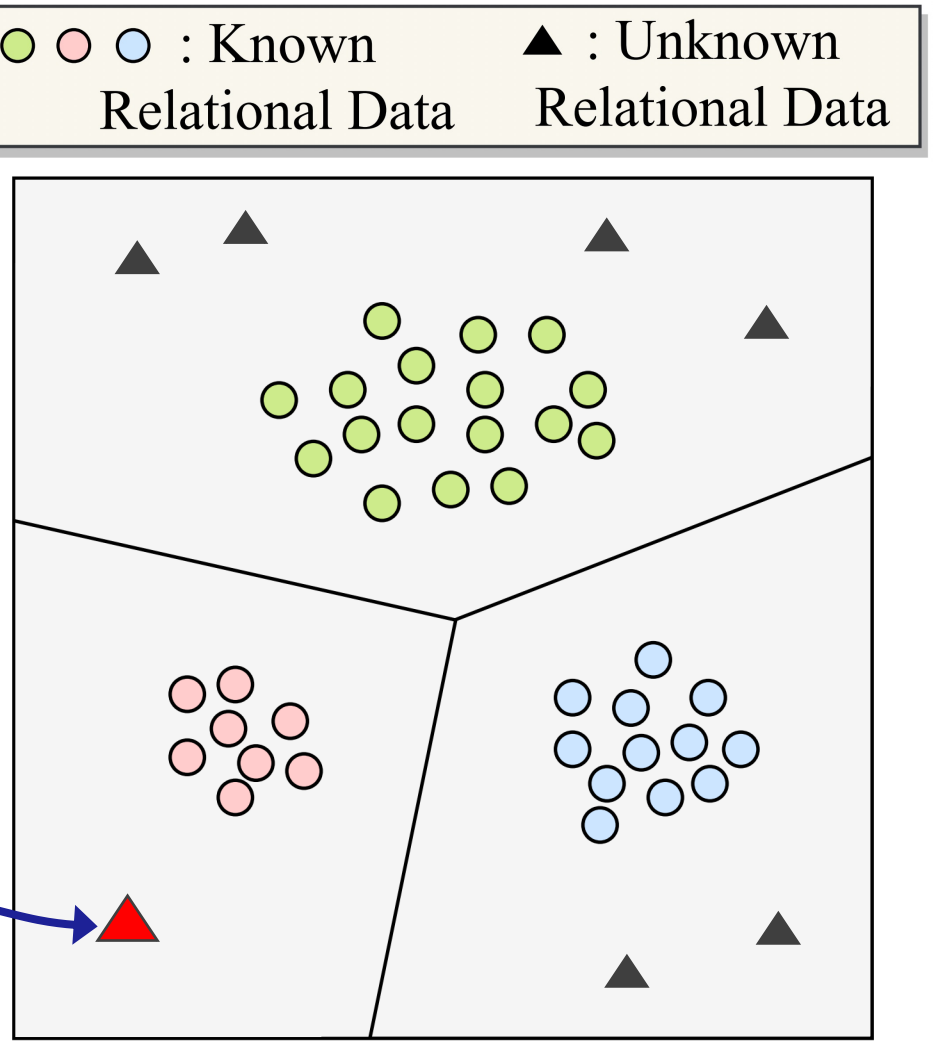
medicines for diseases 0.65

Testing Phase

Carbamazepine may cause intracranial ischemia and cause great harm to patients with cerebral infarction

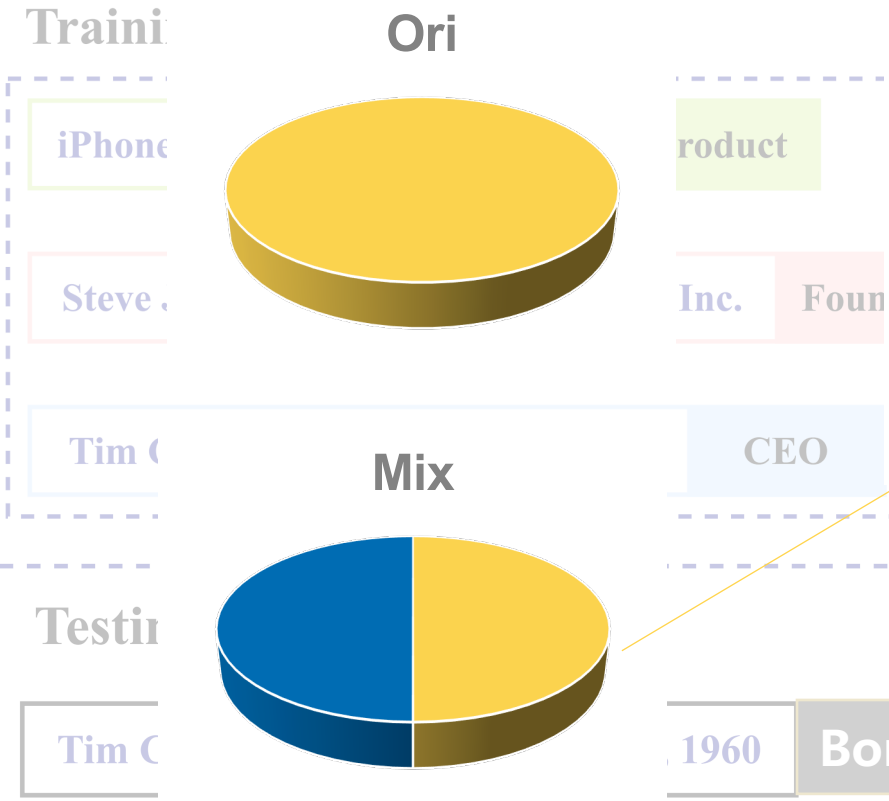
Unknown Relations

contraindicated drugs



Open IE Problems in Knowledge Graph Construction

Closed World Assumption



○ ○ ○ : Known Relational Data

▲ : Unknown Relational Data

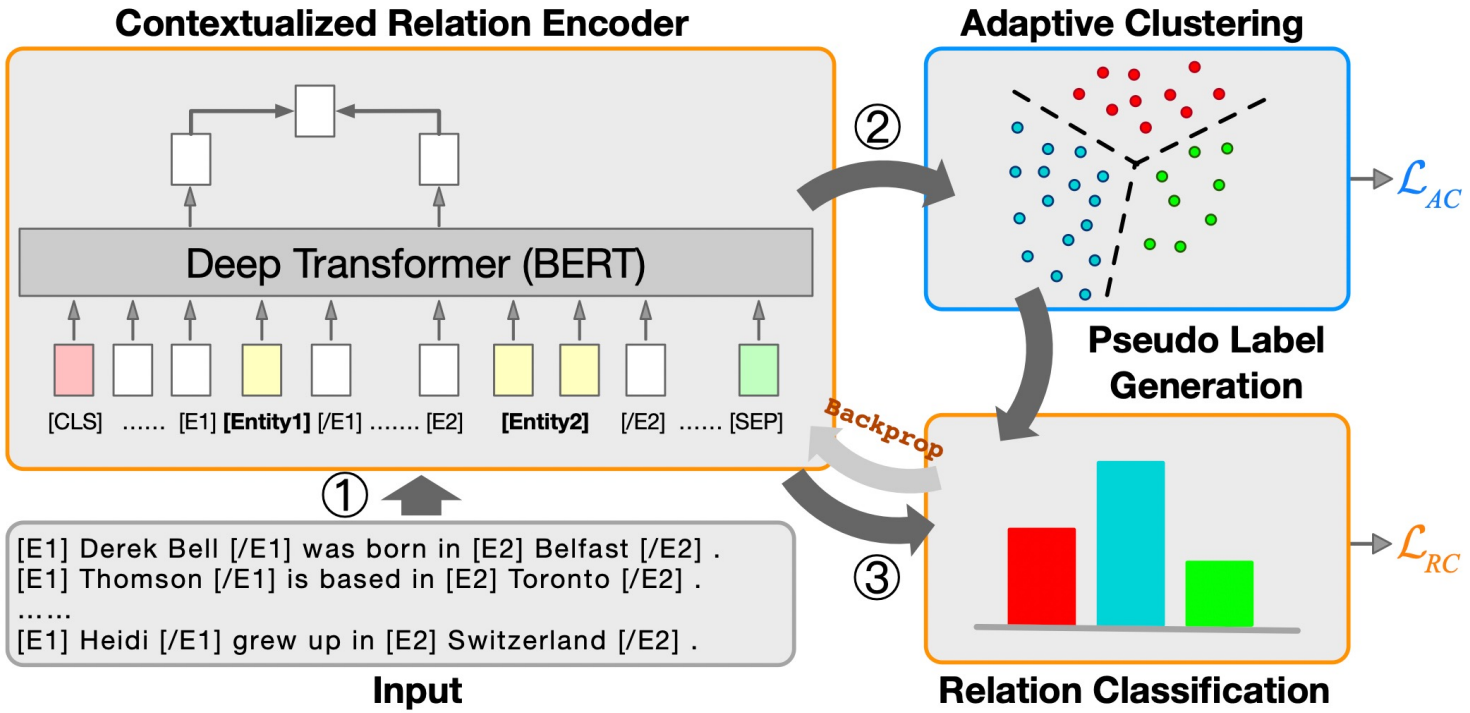
Model/ Dataset	SpanBERT	Roberta	CP
Ori(F_1-score)	0.919	0.928	0.936
Mix(ΔF_1-score)	0.317↓	0.310↓	0.310↓

Unusable in the real world

Open IE Problems in Knowledge Graph Construction

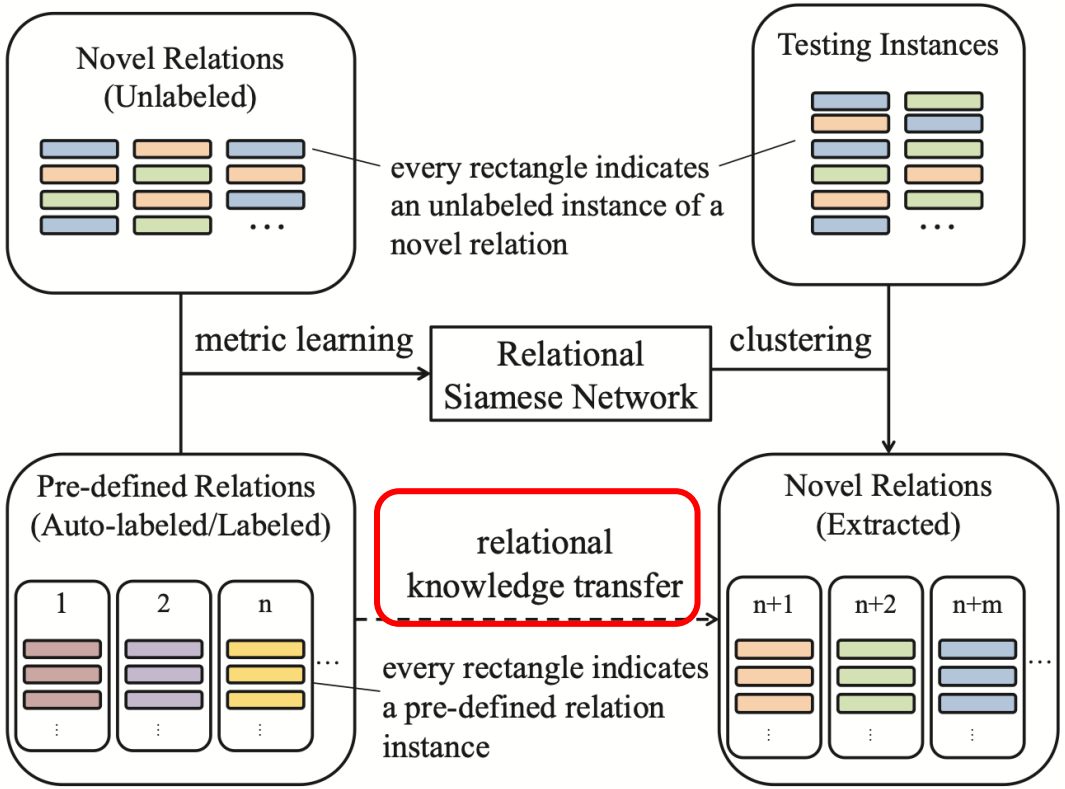
Clustering-based open relation discovery

- Using the relational knowledge encoded in the pretrained model, iteratively optimizes and extracts contextual relational information to cluster relational data

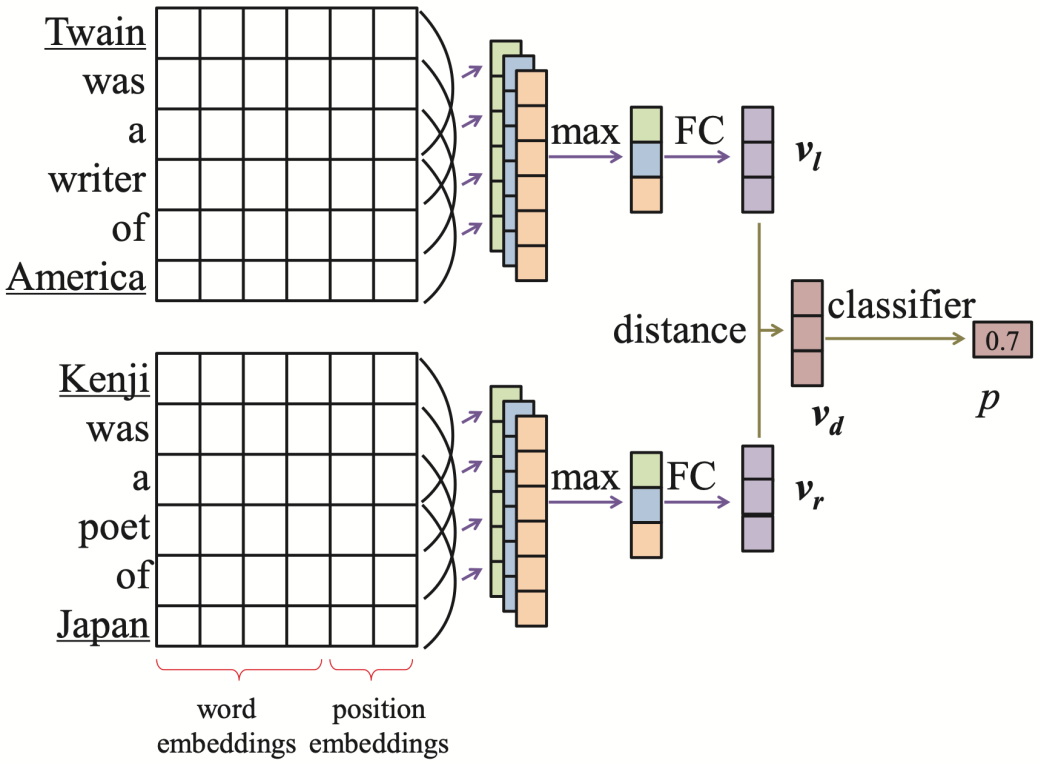


SelfORE: Self-supervised Relational Feature Learning for Open Relation Extraction. EMNLP 2020

Open IE Problems in Knowledge Graph Construction



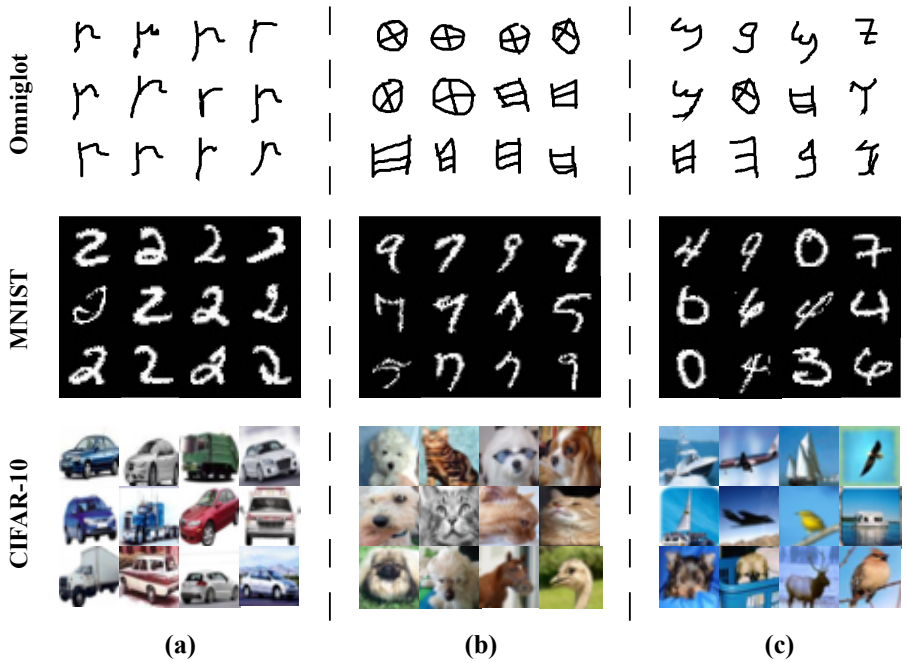
Transfer relational knowledge from predefined relations

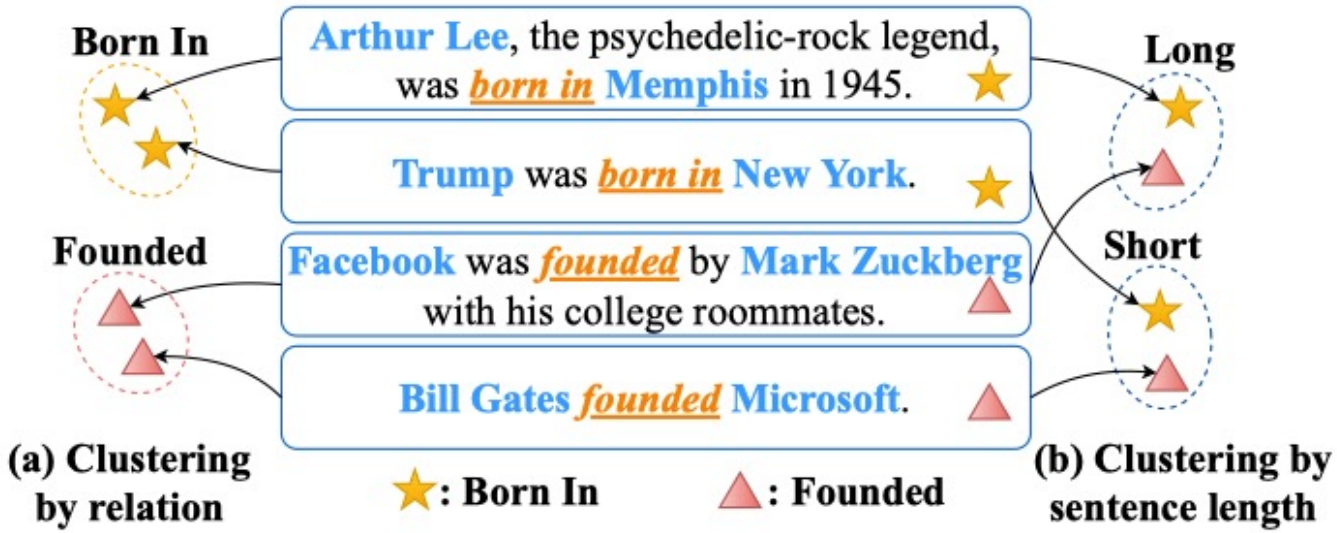


Relational Siamese Network

Wu et al., Open Relation Extraction: Relational Knowledge Transfer from Supervised Data to Unsupervised Data, EMNLP 2019.

Open IE Problems in Knowledge Graph Construction

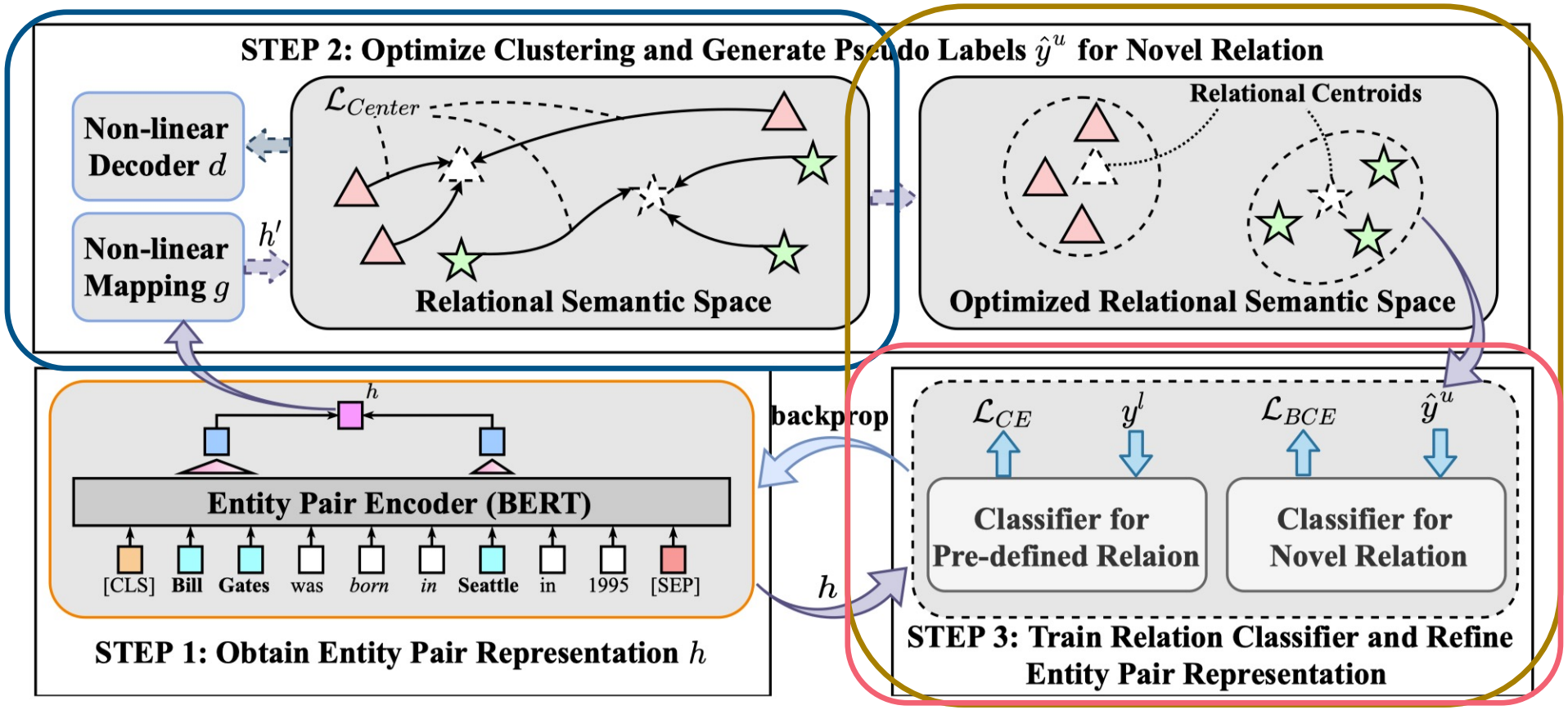




Problems of unsupervised clustering methods

Gui et al., Constructing Multiple Tasks for Augmentation: Improving Neural Image Classification With K-means Features, AAAI 2020.

Open IE Problems in Knowledge Graph Construction



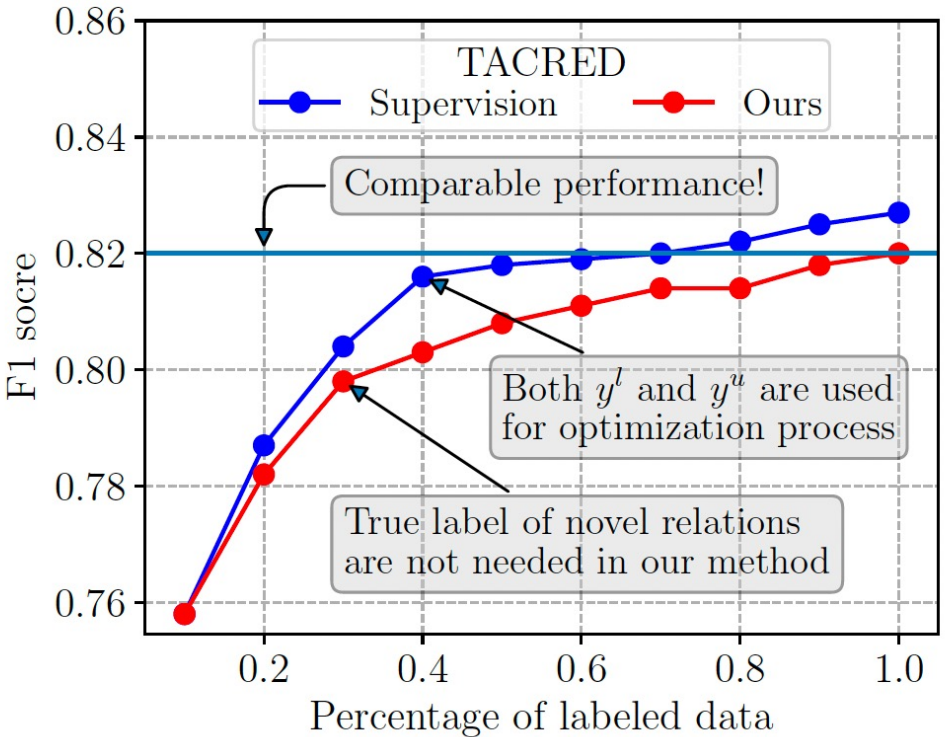
Relation Oriented Clustering Method

A Relation-Oriented Clustering Method for Open Relation Extraction, EMNLP2021

Open IE Problems in Knowledge Graph Construction

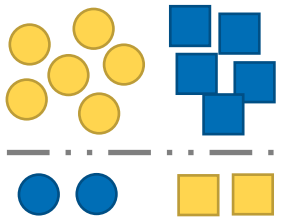
Task	Method	Prec.	Rec.	F_1
	RSN	0.349	0.590	0.439
F → T	RSN-BERT	0.337	0.866	0.486
	RoCORE	0.621_{28}	0.602_{51}	$\mathbf{0.611}_{34}$
	RSN	0.225	0.529	0.316
T → F	RSN-BERT	0.261	0.861	0.400
	RoCORE	0.687_{36}	0.766_{46}	$\mathbf{0.724}_{26}$

Cross Domain Evaluation

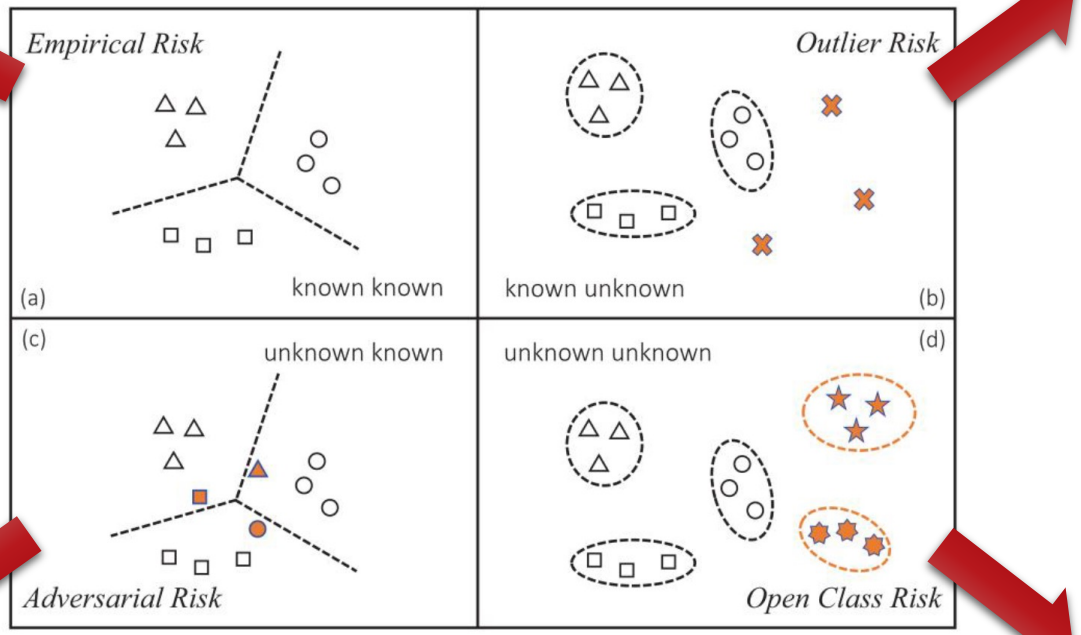


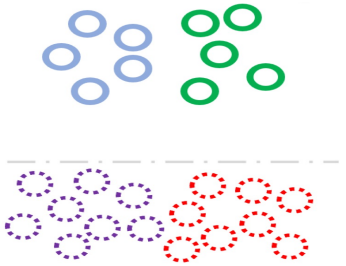
Comparison with labeled data

Robust Knowledge Graph Construction

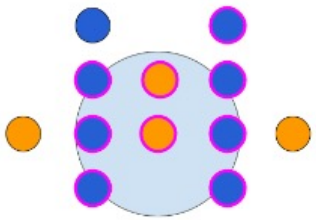


Bias Data

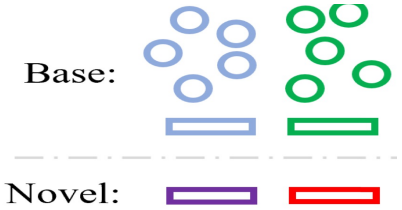




Noisy Data

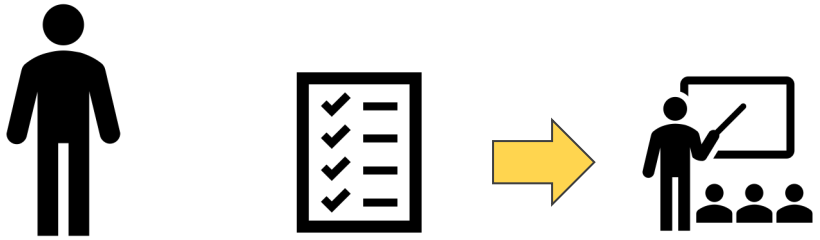


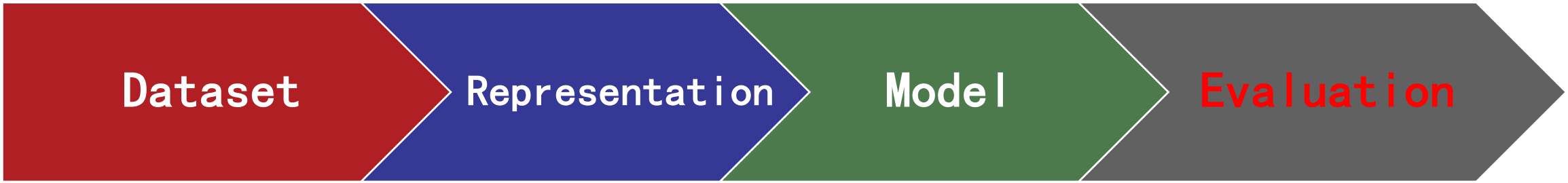
Adversarial Samples



Open IE

Robustness Problems in Knowledge Graph Construction





The system has an impact on the robustness of the model at each step

Robustness evaluation of the model is the **goalkeeper**

Robustness Evaluation – CHECKLIST

CheckList

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Capability	Min Func Test	INVariance	DIRectional
Vocabulary	Fail. rate=15.0%	16.2%	C 34.6%
NER	0.0%	B 20.8%	N/A
Negation	A 76.4%	N/A	N/A
...			

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT***	Labels: negative, positive, neutral		
`Template: I {NEGATION} {POS_VERB} the {THING}.`			
I can't say I recommend the food.	neg	pos	x
I didn't love the flight.	neg	neutral	x
...			
Failure rate = 76.4%			
B Testing **NER** with ***INV***	Same pred. (inv) after removals / additions		
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos / neutral	x
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral / neg	x
...			
Failure rate = 20.8%			
C Testing **Vocabulary** with ***DIR***	Sentiment monotonic decreasing (↓)		
@AmericanAir service wasn't great. You are lame.	↓	neg / neutral	x
@JetBlue why won't YOU help them?! Ugh. I dread you.	↓	neg / neutral	x
...			
Failure rate = 34.6%			

Test NLP models, like we test software

What to test: Linguistic capabilities

How to test: Test behaviors with different test types

Minimum Functionality Test (MFT)

I didn't love the flight.

I can't say I recommend the food.

….

Perturbation tests

INV: Invariance tests

@AmericanAir thank you we got on a different flight to ~~Chicago~~ Dallas.

@VirginAmerica I can't lose my luggage, moving to ~~Brazil~~ Turkey soon

Dir: Directional Expectation Tests

@AmericanAir service wasn't great. You are lame. ↓

@JetBlue why won't YOU help them?! Ugh. I dread you. ↓

Ribeiro et al., *Beyond Accuracy: Behavioral Testing of NLP Models with CheckList*, ACL 2020.

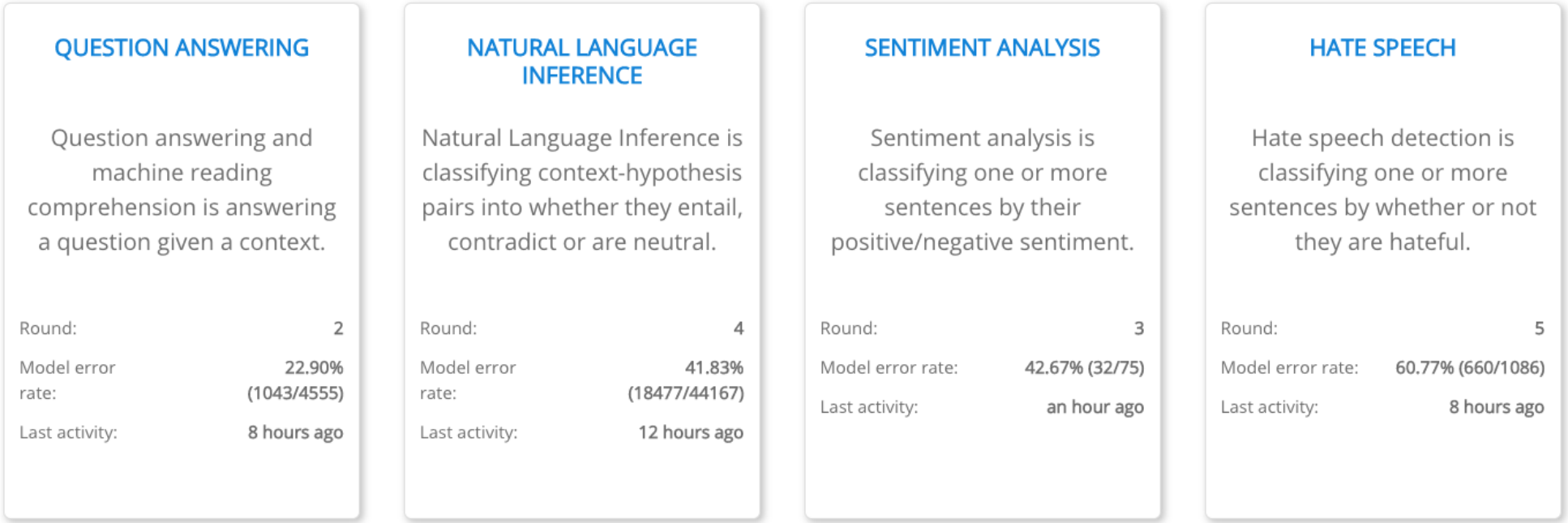
Robustness Evaluation – Dynabench

Dynabench is a research platform for dynamic data collection and benchmarking.

FACEBOOK AI

This platform in essence is a scientific experiment: can we make faster progress if we collect data dynamically, with humans and models in the loop, rather than in the old-fashioned static way?

Dyna Bench



Kiela et al., *Dynabench: Rethinking Benchmarking in NLP*, NAACL 2021.

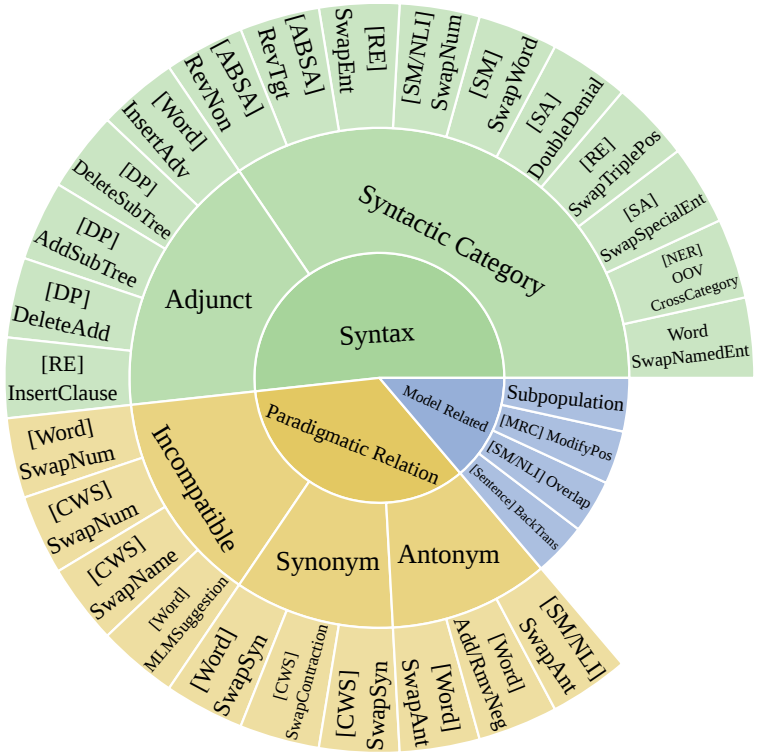
Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing

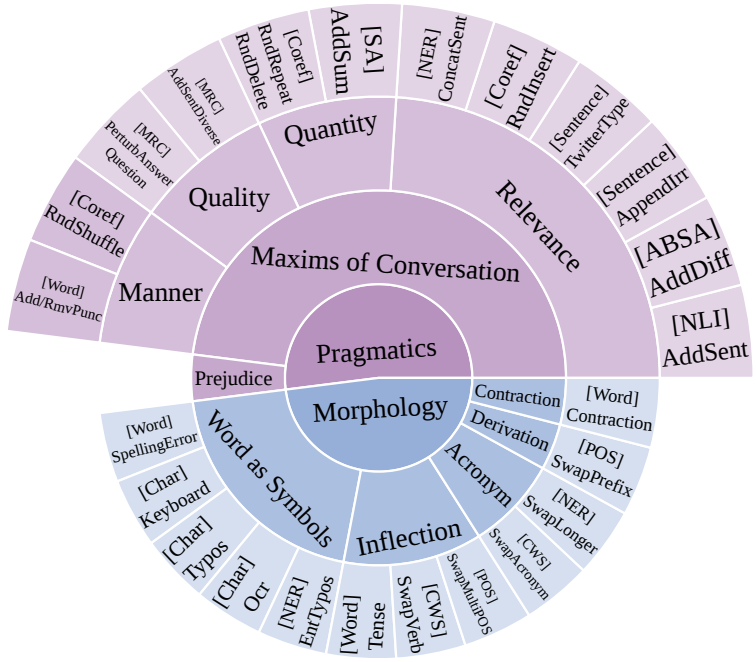
https://github.com/textflint/textflint

4 TextFlint: Robustness Evaluation Toolkit

Integrity

TextFlint offers 20 general transformations, 60 task-specific transformations and thousands of their combinations, and provides over 67,000 evaluation results generated by the transformation on 24 classic datasets from 12 tasks, basically covers all aspects of text transformations to comprehensively evaluate the robustness of a model.





TextFlint

4 TextFlint: Robustness Evaluation Toolkit

Acceptability

Only when the new generated texts conforms to human language, can the robustness result obtained by the verification be credible. Transformation methods provided by TextFlint are scored in plausibility and grammaticality by human evaluation. The results of human and model evaluation can be found on this website.

(a) SA

	Plausibility		Grammaticality	
	Ort.	Trans.	Ort.	Trans.
DoubleDenial	3.26	3.37	3.59	3.49
AddSum-Person	3.39	3.32	3.76	3.59
AddSum-Movie	3.26	3.34	3.61	3.58
SwapSpecialEnt-Person	3.37	3.14	3.75	3.73
SwapSpecialEnt-Movie	3.17	3.28	3.70	3.49

(b) NER

	Plausibility		Grammaticality	
	Ort.	Trans.	Ort.	Trans.
OOV	3.69	3.76	3.54	3.48
SwapLonger	3.73	3.66	3.77	3.54
EntTypos	3.57	3.5	3.59	3.54
CrossCategory	3.48	3.44	3.41	3.32
ConcatSent	4.14	3.54	3.84	3.81

(c) SM

	Plausibility		Grammaticality	
	Ort.	Trans.	Ort.	Trans.
SwapWord	3.08	3.08	3.98	3.92
SwapNum	3.14	3.21	3.87	3.86
Overlap	—	3.33	—	4.11

(d) RE

	Plausibility		Grammaticality	
	Ort.	Trans.	Ort.	Trans.
SwapEnt-MultiType	3.59	3.36	3.97	3.94
SwapEnt-LowFreq	3.34	3.56	3.94	4.05
InsertClause	3.37	3.4	3.89	3.95
SwapEnt-AgeSwap	3.29	3.52	3.85	4.07
SwapTriplePos-BirthSwap	3.52	3.53	3.91	3.86
SwapTriplePos-EmployeeSwap	3.39	3.43	3.88	3.86

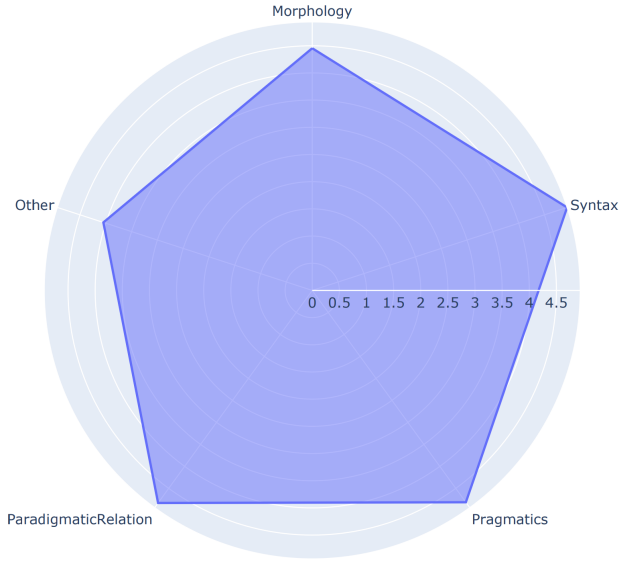
Human evaluation results for task-specific transformation.
Ori. and Trans. represent the original text and the transformed text, respectively.
These metrics are rated on a 1-5 scale (5 for the best).

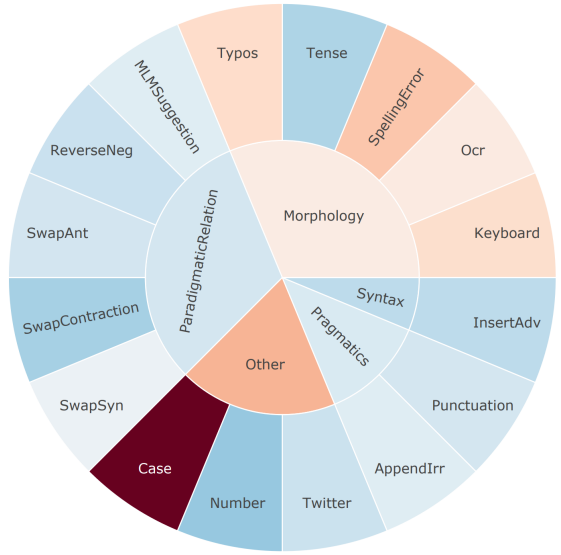
TextFlint

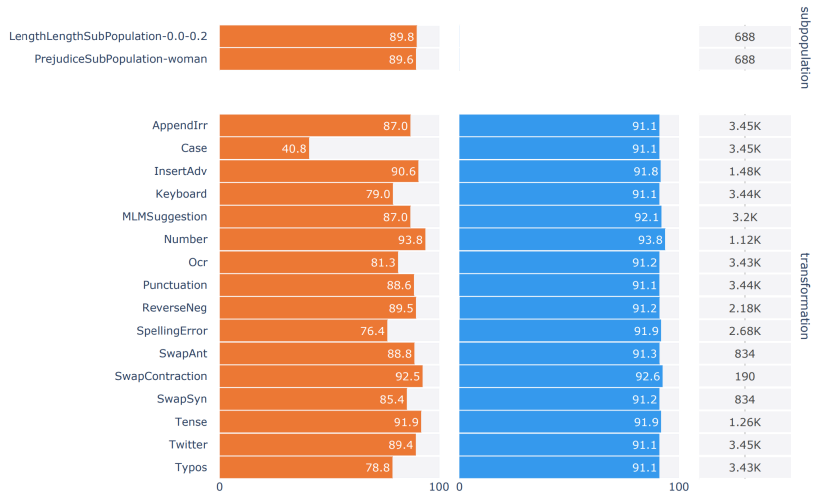
4 TextFlint: Robustness Evaluation Toolkit

Analyzability

TextFlint can give a standard analysis report from the lexics, syntax, semantic levels. All evaluation results can be displayed with visualization and tabulation, so that users can accurately grasp the shortcomings of the model. More evaluation results and related analysis are in the paper.







TextFlint

TestFlint: A Multilingual Robustness Evaluation Toolkit

Table 4: F1 score on the CoNLL 2003 dataset.

Model	*ConcatSent* Ori. → Trans.	*CrossCategory* Ori. → Trans.	*EntTypos* Ori. → Trans.	*OOV* Ori. → Trans.	*SwapLonger* Ori. → Trans.
CoNLL 2003					
CNN-LSTM-CRF (Ma and Hovy, 2016)	90.61 → 87.99	90.59 → 44.18	91.25 → 79.10	90.59 → 58.99	90.59 → 61.15
LSTM-CRF (Lample et al., 2016)	88.49 → 86.88	88.48 → 41.33	89.31 → 74.32	88.48 → 43.55	88.48 → 54.50
LM-LSTM-CRF (Liu et al., 2018)	90.89 → 88.21	90.88 → 44.28	91.54 → 82.90	90.88 → 70.40	90.88 → 65.43
Elmo (Peters et al., 2018)	91.80 → 90.67	91.79 → 44.13	92.48 → 86.19	91.79 → 68.10	91.79 → 61.82
Flair (Akbik et al., 2018)	92.25 → 90.73	92.24 → 45.30	93.05 → 86.78	92.24 → 73.45	92.24 → 66.13
Pooled-Flair (Akbik et al., 2019)	91.90 → 90.45	91.88 → 43.64	92.72 → 86.38	91.88 → 71.70	91.88 → 67.92
TENER (Yan et al., 2019)	91.36 → 90.27	91.35 → 45.43	92.01 → 82.26	91.35 → 55.67	91.35 → 51.10
GRN (Chen et al., 2019)	91.57 → 89.30	91.56 → 42.90	92.29 → 82.72	91.56 → 68.20	91.56 → 65.38
BERT-base (cased) (Devlin et al., 2019)	91.43 → 89.91	91.42 → 44.42	92.20 → 85.02	91.42 → 68.71	91.42 → 79.28
BERT-base (uncased) (Devlin et al., 2019)	90.41 → 90.05	90.40 → 47.19	91.25 → 81.25	90.40 → 64.46	90.40 → 78.26
Average	91.07 → 89.45	91.06 → 44.28	91.81 → 82.69	91.06 → 64.32	91.06 → 65.10

TextFlint

TestFlint: A Multilingual Robustness Evaluation Toolkit

FUDAN UNIVERSITY

English

Aspect-Based Sentiment Analysis
3 domain transformations
16 universal transformations

Coreference Resolution
6 domain transformations
19 universal transformations

Dependency Parsing
2 domain transformations
16 universal transformations

Machine Reading Comprehension
4 domain transformations
13 universal transformations

Named Entity Recognition
5 domain transformations
20 universal transformations

Natural Language Inference
4 domain transformations
17 universal transformations

Part-of-speech tagging
5 domain transformations
19 universal transformations

Relation Extraction
7 domain transformations
15 universal transformations

Sentiment Analysis
5 domain transformations
18 universal transformations

Semantic Matching
3 domain transformations
22 universal transformations

Text Classification
0 domain transformations
18 universal transformations

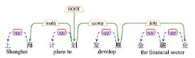
The Winograd Schema Challenge
4 domain transformations
8 universal transformations

Word Sense Disambiguation
1 domain transformations
18 universal transformations

Neural Machine Translation
3 domain transformations
11 universal transformations

Chinese

Chinese Word Segmentation
7 domain transformations
0 universal transformations



Chinese Dependency Parsing
0 domain transformations
8 universal transformations

Chinese Named Entity Recognition
4 domain transformations
5 universal transformations

Chinese Semantic Matching
3 domain transformations
12 universal transformations

Chinese Machine Reading Comprehension
3 domain transformations
15 universal transformations

Chinese Sentiment Analysis
3 domain transformations
8 universal transformations

TextFlint

Architecture