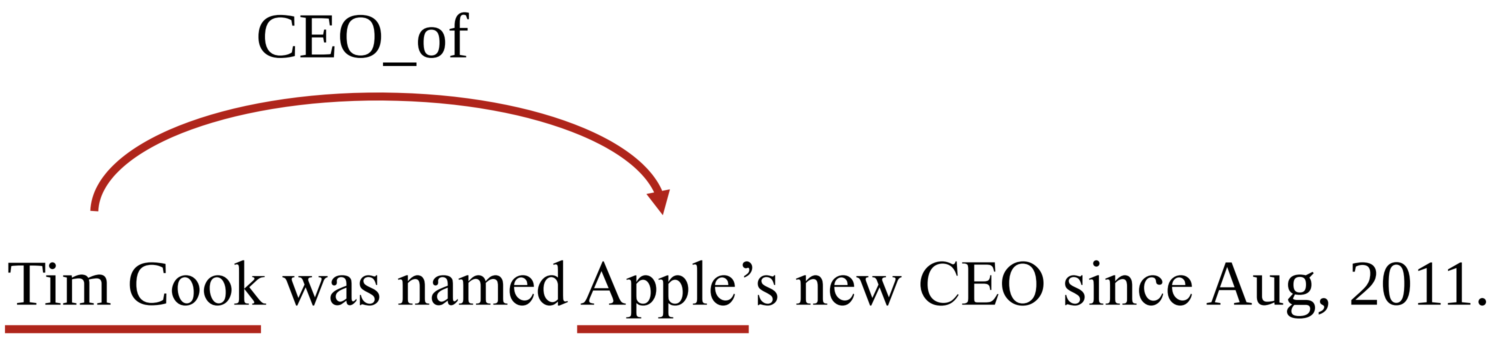
Towards Accurate and Consistent Evaluation:
A Dataset for Distantly-Supervised Relation Extraction

Tong Zhu, Haitao Wang, Junjie Yu, Xiabing Zhou,
Wenliang Chen, Wei Zhang, Min Zhang

Institute of Artificial Intelligence, School of Computer Science and Technology,
Soochow University, China
Alibaba Group, China

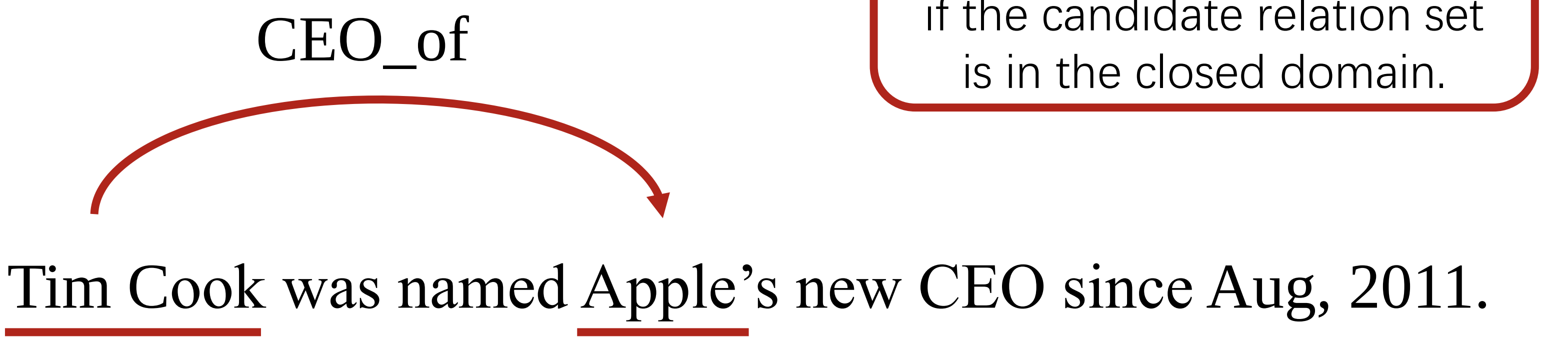
Relation Extraction (RE)



CEO_of

Tim Cook was named Apple’s new CEO since Aug, 2011.

Relation Extraction (RE)



CEO_of

Tim Cook was named Apple’s new CEO since Aug, 2011.

RE can be regarded as a **classification task** if the candidate relation set is in the closed domain.

Distantly-Supervised Relation Extraction (DSRE)

- Basic Distant Supervision (DS) assumption:

If two entities participate in a relation, **all sentences** that mention these two entities express that relation.

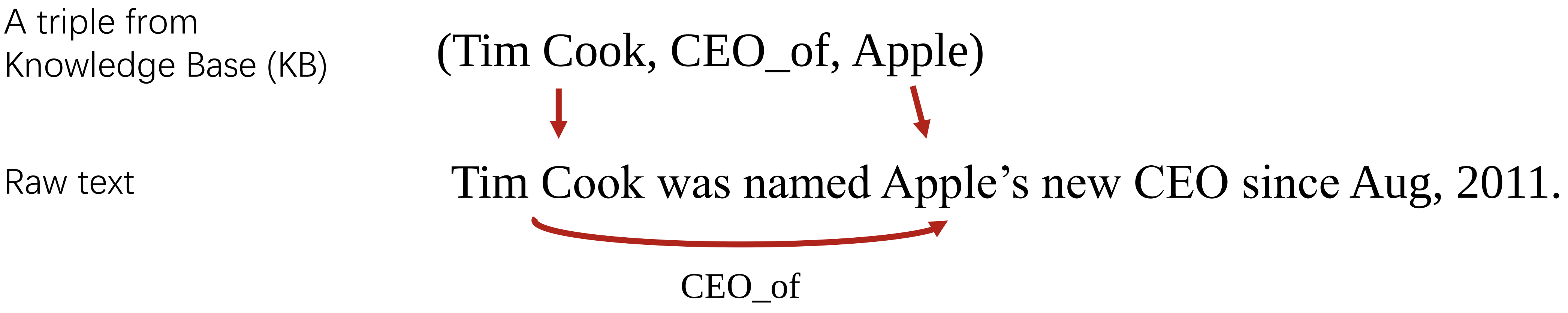
M. Mintz, S. Bills, R. Snow, and D. Jurafsky, “Distant supervision for relation extraction without labeled data,” in *ACL*, 2009

Distantly-Supervised Relation Extraction (DSRE)

- Basic Distant Supervision (DS) assumption:

> If two entities participate in a relation, **all sentences** that mention these two entities express that relation.

- Example:



A triple from Knowledge Base (KB)

(Tim Cook, CEO_of, Apple)

Raw text

Tim Cook was named Apple’s new CEO since Aug, 2011.

CEO_of

M. Mintz, S. Bills, R. Snow, and D. Jurafsky, “Distant supervision for relation extraction without labeled data,” in *ACL*, 2009

Problem 1 – DS Noises

Knowledge Base		
Head	**Tail**	**Relation**
Apple	Steve Jobs	founders
Adele	London	place_of_birth
Aragaki Yui	Japanese	nationality

Corpus		
Sentence	**Distantly-Supervised Label**	**Comment**
Steve Jobs left ***Apple*** in 1985 .	founders	✗ False Positive
Adele was born in the ***UK*** .	NA	✗ False Negative
Aragaki Yui is an ***Japanese*** actress .	nationality	✓ True Positive
Jack, have you heard of ***Hemingway*** ?	NA	✓ True Negative

Problem 1 – DS Noises

Knowledge Base		
Head	**Tail**	**Relation**
Apple	Steve Jobs	founders
Adele	London	place_of_birth
Aragaki Yui	Japanese	nationality

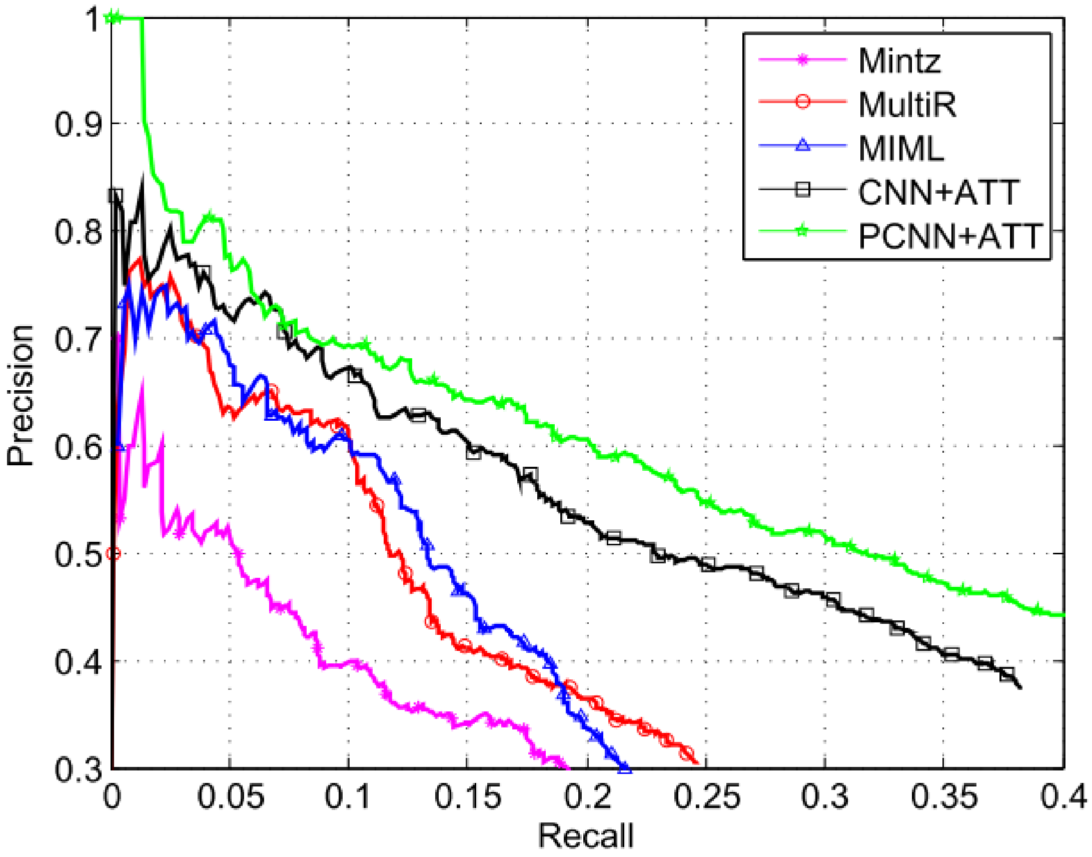
Corpus		
Sentence	**Distantly-Supervised Label**	**Comment**
Steve Jobs left ***Apple*** in 1985 .	founders	✗ False Positive
Adele was born in the ***UK*** .	NA	✗ False Negative
Aragaki Yui is an ***Japanese*** actress .	nationality	✓ True Positive
Jack, have you heard of ***Hemingway*** ?	NA	✓ True Negative

Problem 1 – DS Noises

Knowledge Base		
Head	**Tail**	**Relation**
Apple	Steve Jobs	founders
Adele	London	place_of_birth
Aragaki Yui	Japanese	nationality

Corpus		
Sentence	**Distantly-Supervised Label**	**Comment**
Steve Jobs left ***Apple*** in 1985 .	founders	✗ False Positive
Adele was born in the ***UK*** .	NA	✗ False Negative
Aragaki Yui is an ***Japanese*** actress .	nationality	✓ True Positive
Jack, have you heard of ***Hemingway*** ?	NA	✓ True Negative

Problem 2 – Evaluation Metrics Bias



Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, “Neural relation extraction with selective attention over instances,” in *ACL*, 2016, vol. 4, pp. 2124–2133

- Precision Recall Curve (PRC) & Area Under Curve (AUC)

Solutions

- Solution 1 – Noise Deduction while Training
  - Multi-Instance Learning (MIL) groups texts with the same entity pairs as one bag, and train & test on bag-level
  - MIL follows a weaker assumption of DS

- Solution 2 – Manually Check the Predictions
  - PRC & AUC are not precisely reliable metrics for evaluation
  - Precision@K scores are the final criteria with human annotations

Solutions and Remained Problems

- Solution 1 – Noise Deduction while Training
  - Multi-Instance Learning (MIL) groups texts with the same entity pairs as one bag, and train & test on bag-level
  - MIL follows a weaker assumption of DS

MIL only mitigates the noise effects during training

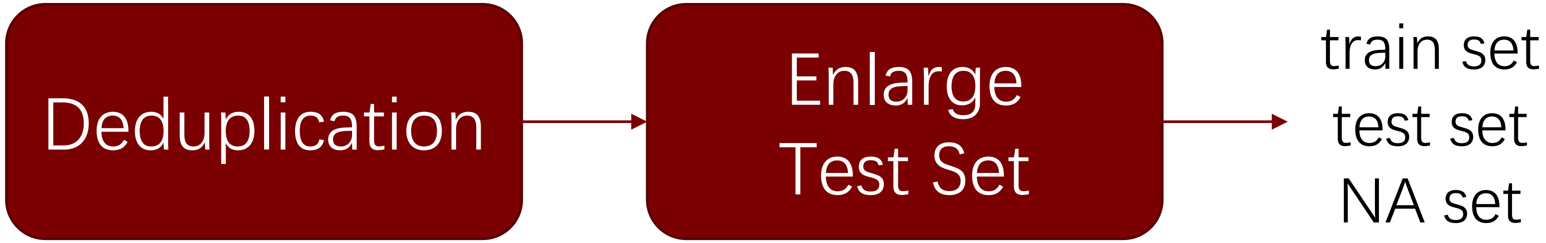
- Solution 2 – Manually Check the Predictions
  - PRC & AUC are not precisely reliable metrics for evaluation
  - Precision@K scores are the final criteria with human annotations

AUC scores are not reliable due to the noises in the test set

a. Different papers have different annotation criteria
b. K value is usually too small to cover all the relations

Dataset Construction – Preprocessing

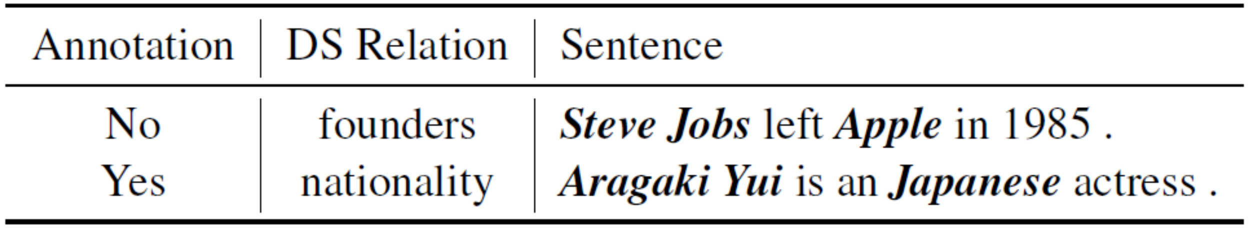
- NYT-H is built on NYT10



S. Riedel, L. Yao, and A. McCallum, "Modeling Relations and Their Mentions without Labeled Text BT - Machine Learning and Knowledge Discovery in Databases," in *ECML PKDD*, 2010, pp. 148–163.

Dataset Construction – Annotation

- There are over 50 relation types in original NYT10
- A binary strategy is applied to ease the annotation task
- 10,065 sentences are labelled with a Kappa coefficient of 0.753



Annotation	DS Relation	Sentence
No	founders	***Steve Jobs*** left ***Apple*** in 1985 .
Yes	nationality	***Aragaki Yui*** is an ***Japanese*** actress .

Annotation Example

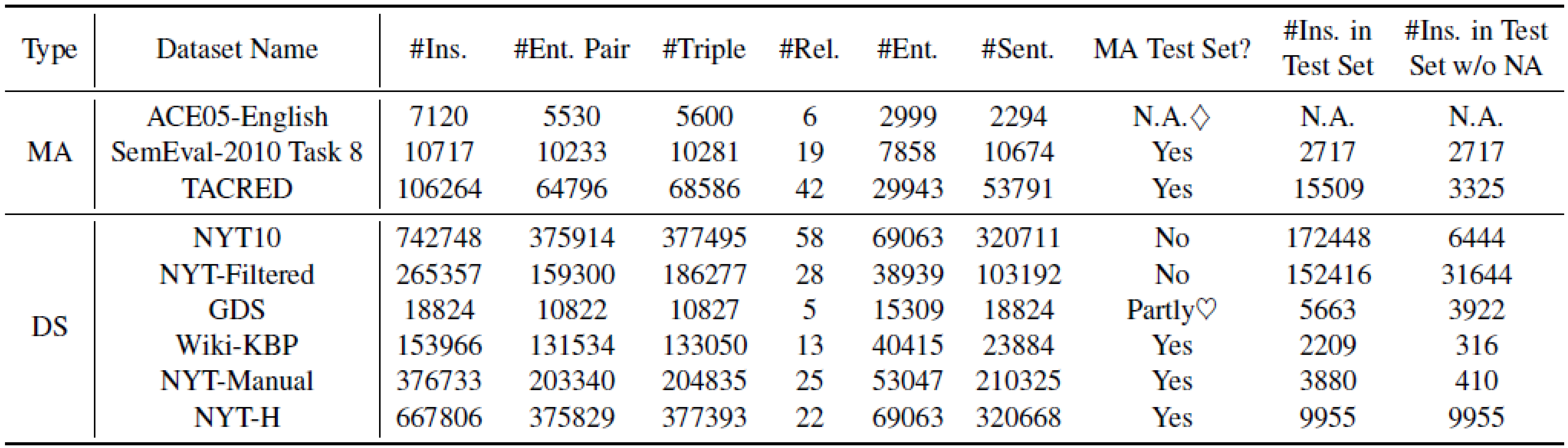
Dataset Construction – Postprocessing

- The following relations are converted into NA relation
  - Relations that does not occur both in the train and test set
  - If the number of instances are less than 100 in the train set
  - If there are no instances labelled as “Yes” in the test set
- 22 relations (including NA) are kept in the final version of NYT-H

Dataset	#Instance	#Bag	#Yes Instance	#Yes Bag
NA	550,720	357,196	/	/
Train	107,093	16,370	/	/
Test	9,955	3,548	5,202	2,277

Data Statistics

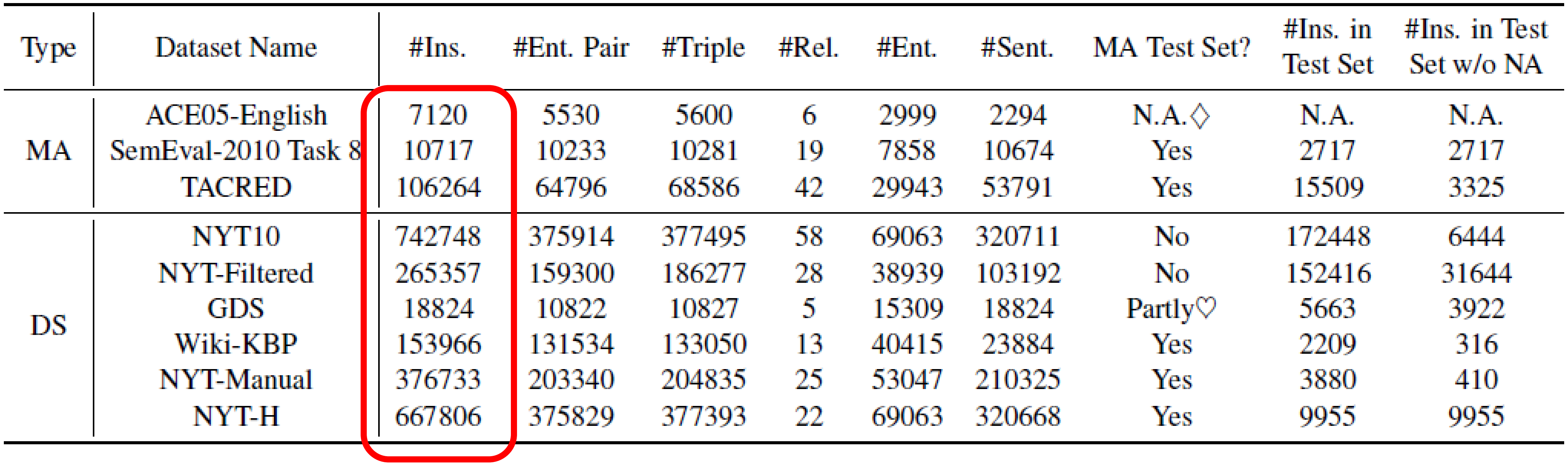
Dataset Comparisons



Type	Dataset Name	#Ins.	#Ent. Pair	#Triple	#Rel.	#Ent.	#Sent.	MA Test Set?	#Ins. in Test Set	#Ins. in Test Set w/o NA
MA	ACE05-English	7120	5530	5600	6	2999	2294	N.A.♢	N.A.	N.A.
	SemEval-2010 Task 8	10717	10233	10281	19	7858	10674	Yes	2717	2717
	TACRED	106264	64796	68586	42	29943	53791	Yes	15509	3325
DS	NYT10	742748	375914	377495	58	69063	320711	No	172448	6444
	NYT-Filtered	265357	159300	186277	28	38939	103192	No	152416	31644
	GDS	18824	10822	10827	5	15309	18824	Partly♡	5663	3922
	Wiki-KBP	153966	131534	133050	13	40415	23884	Yes	2209	316
	NYT-Manual	376733	203340	204835	25	53047	210325	Yes	3880	410
	NYT-H	667806	375829	377393	22	69063	320668	Yes	9955	9955

Dataset Comparisons: MA: fully manually annotated

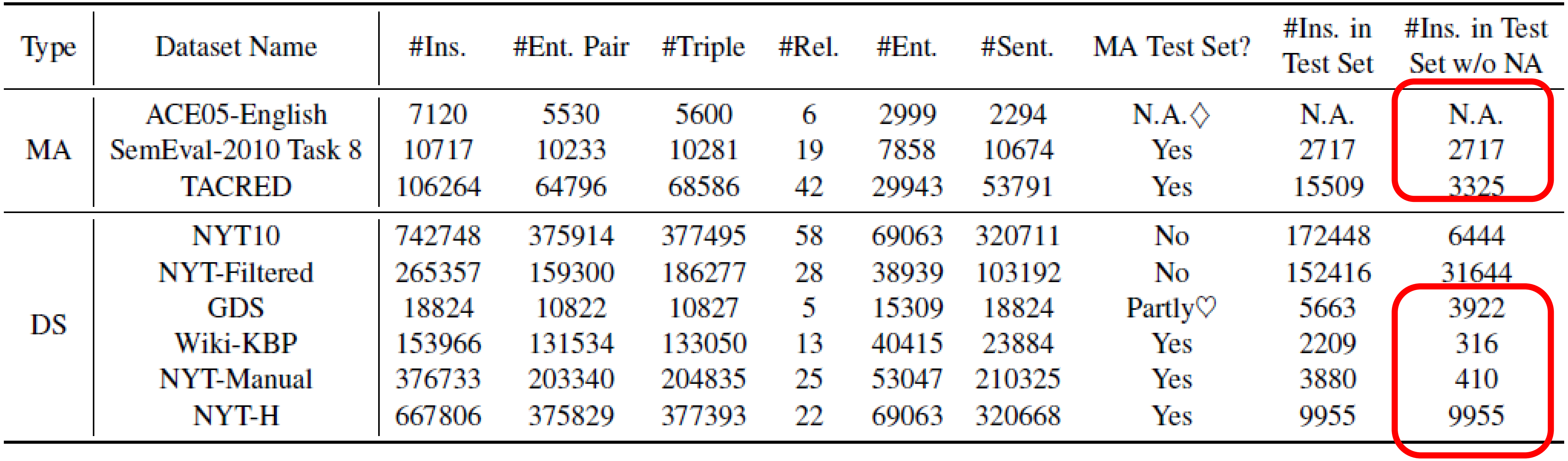
Dataset Comparisons



Type	Dataset Name	#Ins.	#Ent. Pair	#Triple	#Rel.	#Ent.	#Sent.	MA Test Set?	#Ins. in Test Set	#Ins. in Test Set w/o NA
MA	ACE05-English	7120	5530	5600	6	2999	2294	N.A.♢	N.A.	N.A.
	SemEval-2010 Task 8	10717	10233	10281	19	7858	10674	Yes	2717	2717
	TACRED	106264	64796	68586	42	29943	53791	Yes	15509	3325
DS	NYT10	742748	375914	377495	58	69063	320711	No	172448	6444
	NYT-Filtered	265357	159300	186277	28	38939	103192	No	152416	31644
	GDS	18824	10822	10827	5	15309	18824	Partly♡	5663	3922
	Wiki-KBP	153966	131534	133050	13	40415	23884	Yes	2209	316
	NYT-Manual	376733	203340	204835	25	53047	210325	Yes	3880	410
	NYT-H	667806	375829	377393	22	69063	320668	Yes	9955	9955

Dataset Comparisons: MA: fully manually annotated

Dataset Comparisons



Type	Dataset Name	#Ins.	#Ent. Pair	#Triple	#Rel.	#Ent.	#Sent.	MA Test Set?	#Ins. in Test Set	#Ins. in Test Set w/o NA
MA	ACE05-English	7120	5530	5600	6	2999	2294	N.A.♢	N.A.	N.A.
	SemEval-2010 Task 8	10717	10233	10281	19	7858	10674	Yes	2717	2717
	TACRED	106264	64796	68586	42	29943	53791	Yes	15509	3325
DS	NYT10	742748	375914	377495	58	69063	320711	No	172448	6444
	NYT-Filtered	265357	159300	186277	28	38939	103192	No	152416	31644
	GDS	18824	10822	10827	5	15309	18824	Partly♡	5663	3922
	Wiki-KBP	153966	131534	133050	13	40415	23884	Yes	2209	316
	NYT-Manual	376733	203340	204835	25	53047	210325	Yes	3880	410
	NYT-H	667806	375829	377393	22	69063	320668	Yes	9955	9955

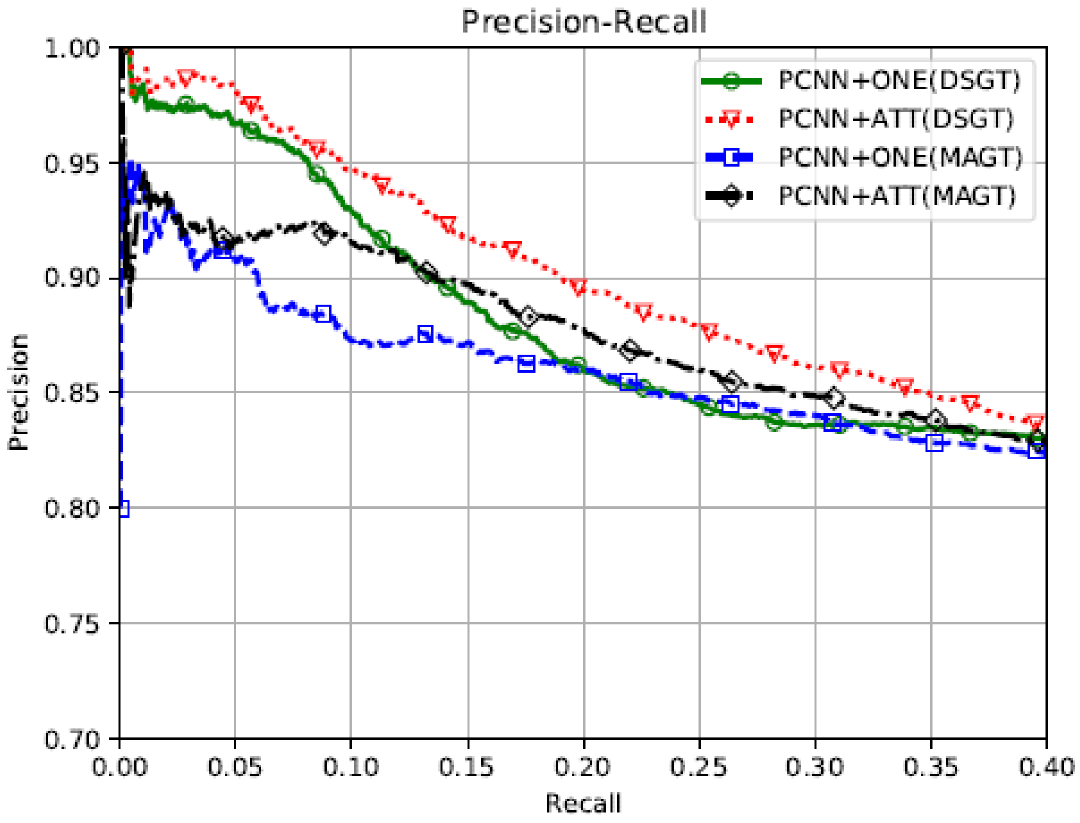
Dataset Comparisons: MA: fully manually annotated

Evaluation Tracks & Measures

- Track
  - Sent2Sent: Train at sentence-level and evaluate at sentence-level
  - Bag2Sent: Train at bag-level and evaluate at sentence-level
  - Bag2Bag: Train at bag-level and evaluate at bag-level

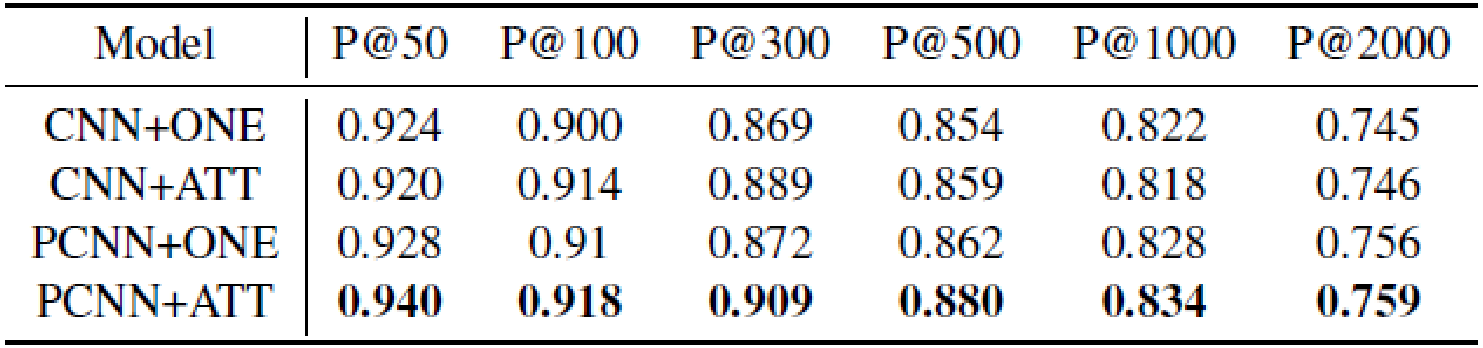
- Measure
  - DSGT: **D**istantly-**S**upervised relation as **G**round **T**ruth
  - MAGT: **M**anually **A**nnotated relation as **G**round **T**ruth

PRC Results in Bag2Bag Track



PRC in Bag2Bag Track

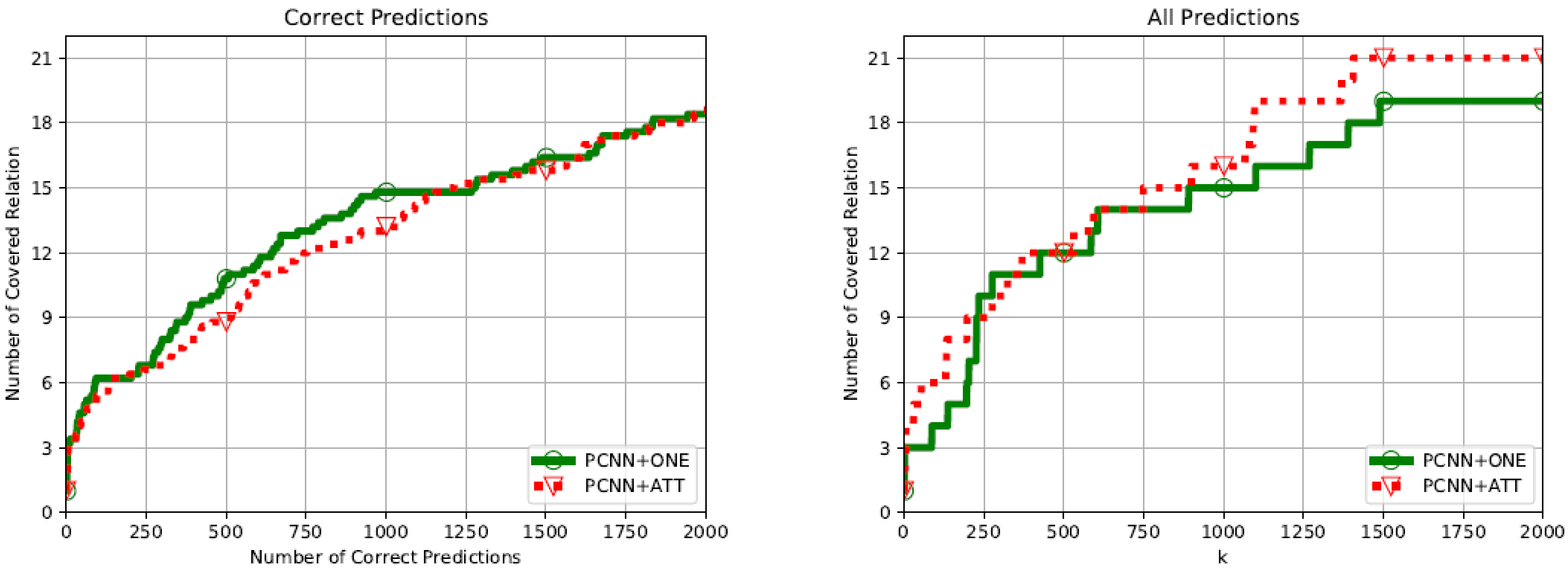
Precision@K Results of Bag2Bag Track



Model	P@50	P@100	P@300	P@500	P@1000	P@2000
CNN+ONE	0.924	0.900	0.869	0.854	0.822	0.745
CNN+ATT	0.920	0.914	0.889	0.859	0.818	0.746
PCNN+ONE	0.928	0.91	0.872	0.862	0.828	0.756
PCNN+ATT	**0.940**	**0.918**	**0.909**	**0.880**	**0.834**	**0.759**

Precision@K Results in Bag2Bag Track

Relation Coverage in Precision@K Evaluation

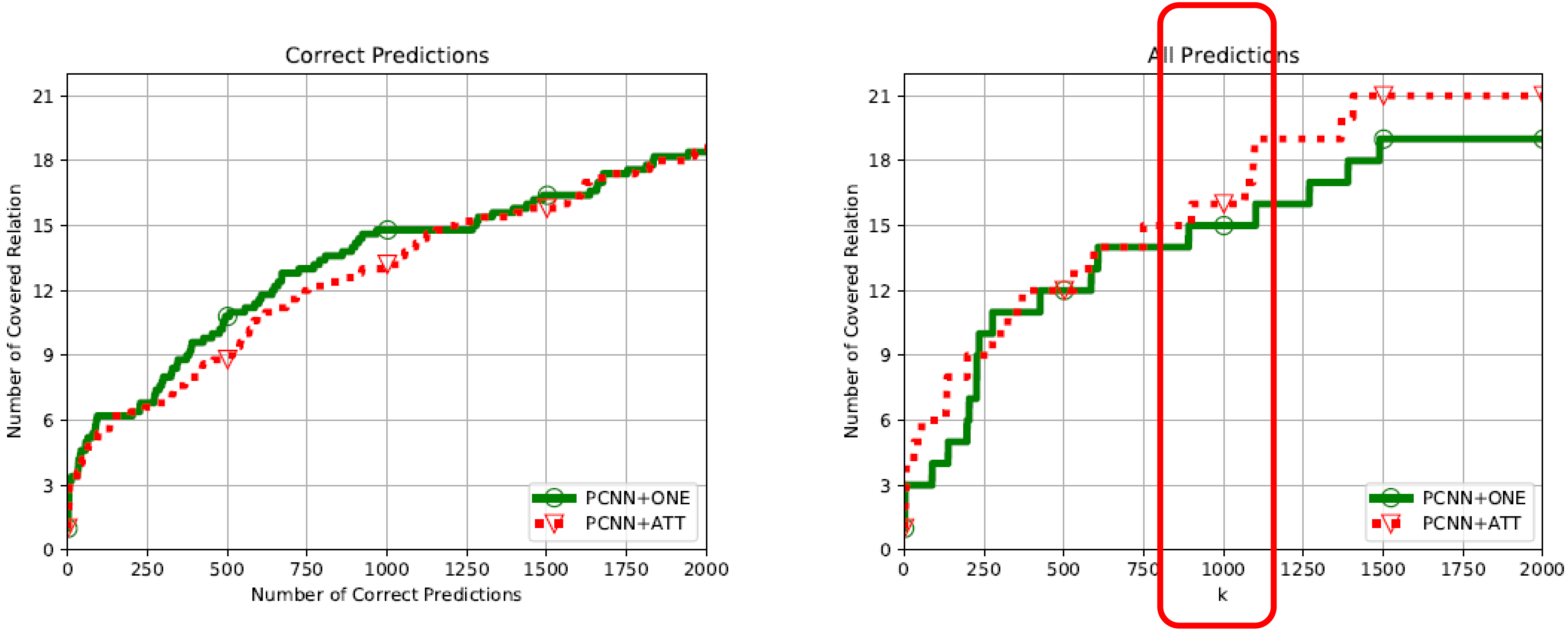


(a) Relation Coverage for Correct Predictions

(b) Relation Coverage for All Predictions

Relation Coverage

Relation Coverage in Precision@K Evaluation

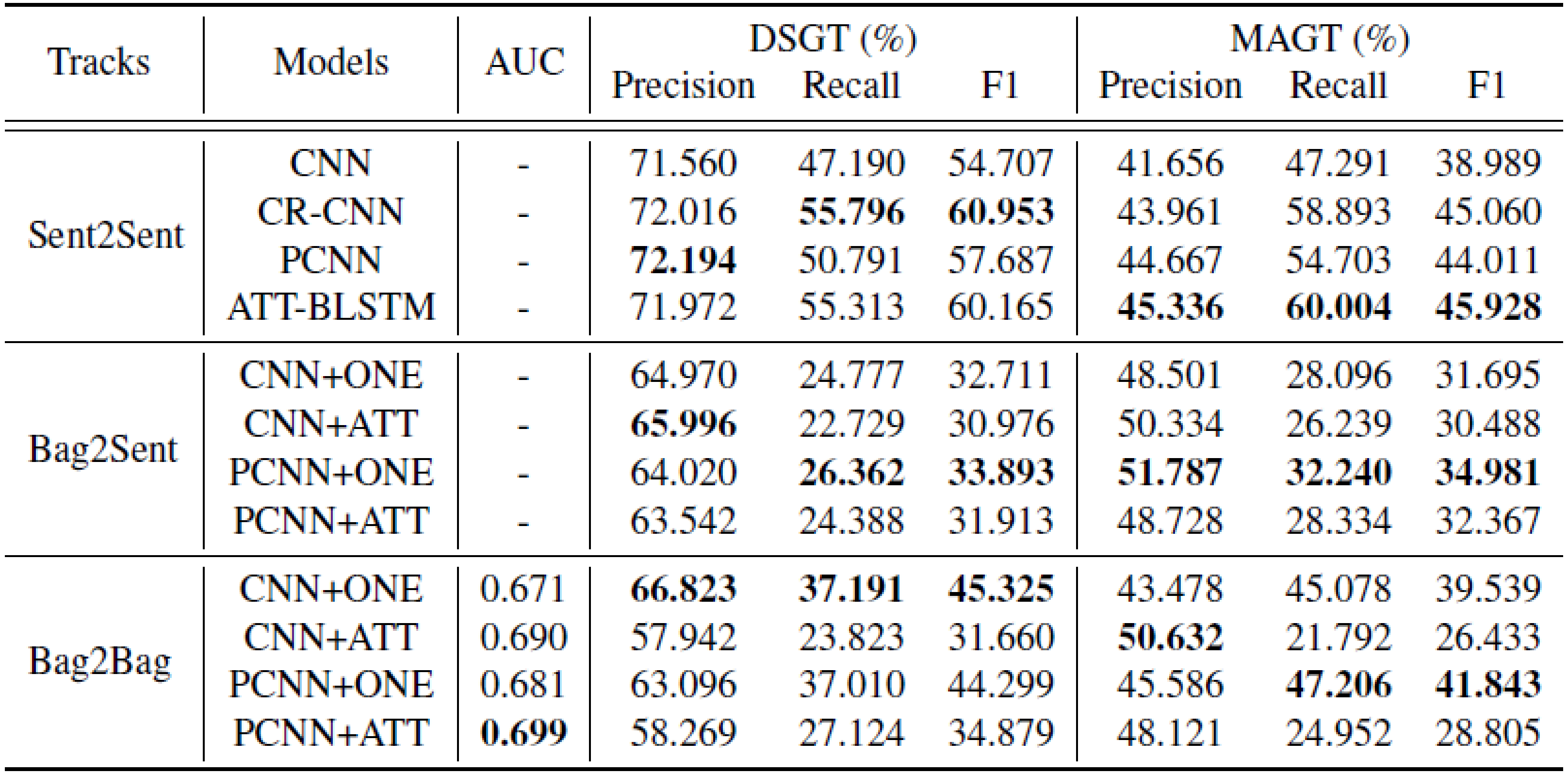


(a) Relation Coverage for Correct Predictions

(b) Relation Coverage for All Predictions

Relation Coverage

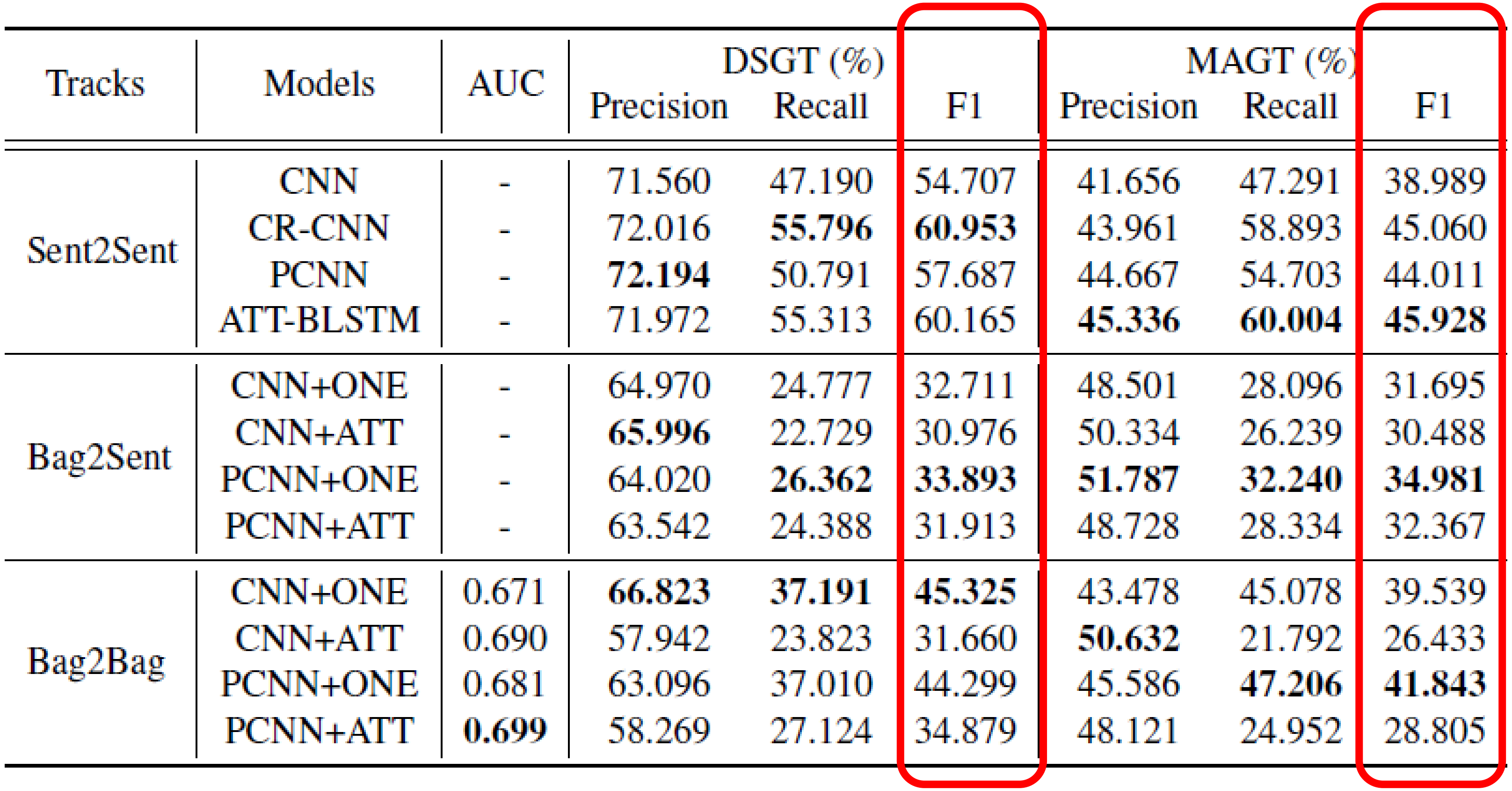
AUC & F1 Results



Tracks	Models	AUC	DSGT (%)			MAGT (%)		
			Precision	Recall	F1	Precision	Recall	F1
Sent2Sent	CNN	-	71.560	47.190	54.707	41.656	47.291	38.989
	CR-CNN	-	72.016	**55.796**	**60.953**	43.961	58.893	45.060
	PCNN	-	**72.194**	50.791	57.687	44.667	54.703	44.011
	ATT-BLSTM	-	71.972	55.313	60.165	**45.336**	**60.004**	**45.928**
Bag2Sent	CNN+ONE	-	64.970	24.777	32.711	48.501	28.096	31.695
	CNN+ATT	-	**65.996**	22.729	30.976	50.334	26.239	30.488
	PCNN+ONE	-	64.020	**26.362**	**33.893**	**51.787**	**32.240**	**34.981**
	PCNN+ATT	-	63.542	24.388	31.913	48.728	28.334	32.367
Bag2Bag	CNN+ONE	0.671	**66.823**	**37.191**	**45.325**	43.478	45.078	39.539
	CNN+ATT	0.690	57.942	23.823	31.660	**50.632**	21.792	26.433
	PCNN+ONE	0.681	63.096	37.010	44.299	45.586	**47.206**	**41.843**
	PCNN+ATT	**0.699**	58.269	27.124	34.879	48.121	24.952	28.805

AUC & macro-F1 Scores among All Tracks

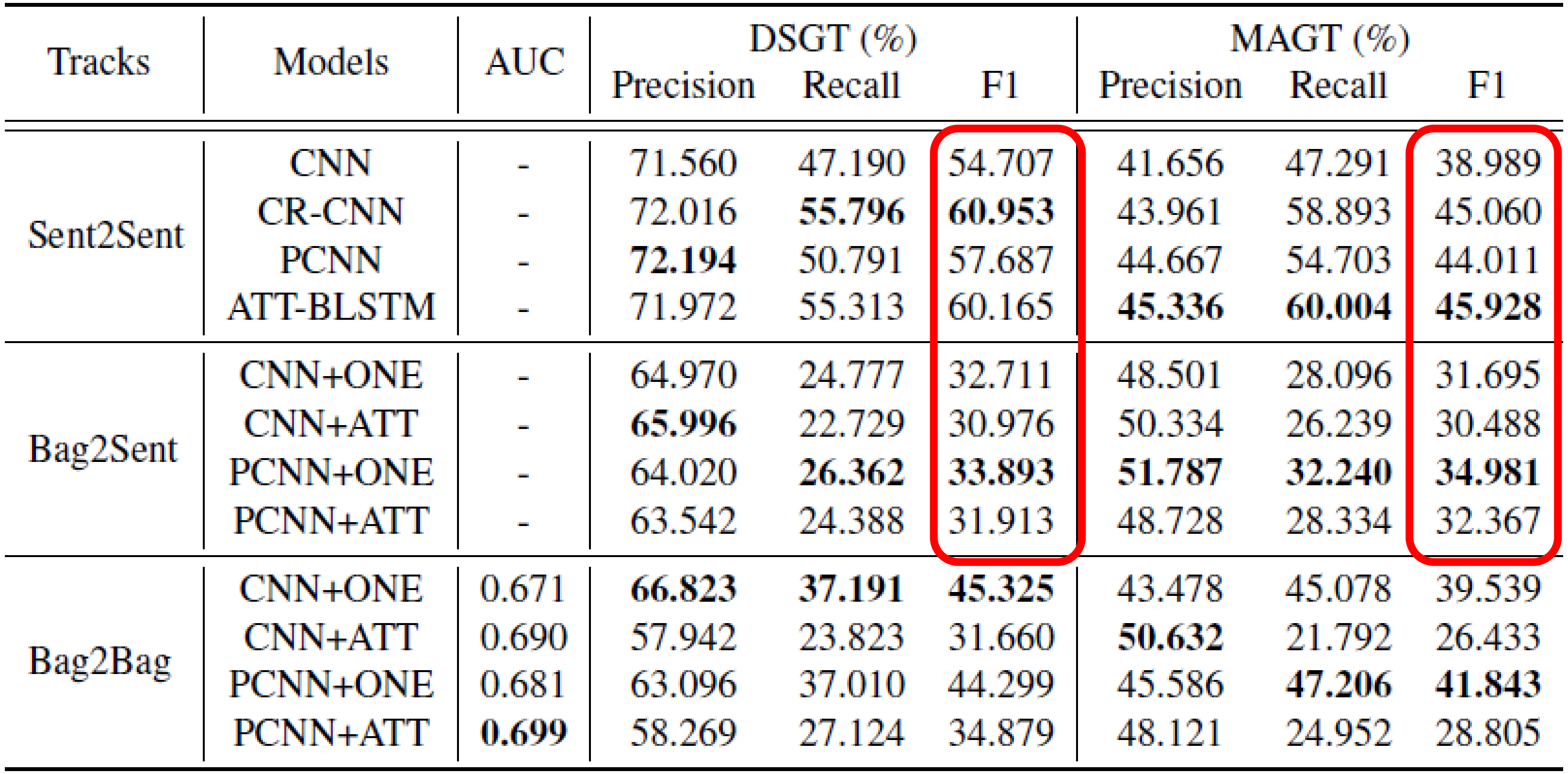
AUC & F1 Results



Tracks	Models	AUC	DSGT (%)			MAGT (%)		
			Precision	Recall	F1	Precision	Recall	F1
Sent2Sent	CNN	-	71.560	47.190	54.707	41.656	47.291	38.989
	CR-CNN	-	72.016	**55.796**	**60.953**	43.961	58.893	45.060
	PCNN	-	**72.194**	50.791	57.687	44.667	54.703	44.011
	ATT-BLSTM	-	71.972	55.313	60.165	**45.336**	**60.004**	**45.928**
Bag2Sent	CNN+ONE	-	64.970	24.777	32.711	48.501	28.096	31.695
	CNN+ATT	-	**65.996**	22.729	30.976	50.334	26.239	30.488
	PCNN+ONE	-	64.020	**26.362**	**33.893**	**51.787**	**32.240**	**34.981**
	PCNN+ATT	-	63.542	24.388	31.913	48.728	28.334	32.367
Bag2Bag	CNN+ONE	0.671	**66.823**	**37.191**	**45.325**	43.478	45.078	39.539
	CNN+ATT	0.690	57.942	23.823	31.660	**50.632**	21.792	26.433
	PCNN+ONE	0.681	63.096	37.010	44.299	45.586	**47.206**	**41.843**
	PCNN+ATT	**0.699**	58.269	27.124	34.879	48.121	24.952	28.805

AUC & macro-F1 Scores among All Tracks

AUC & F1 Results



Tracks	Models	AUC	DSGT (%)			MAGT (%)		
			Precision	Recall	F1	Precision	Recall	F1
Sent2Sent	CNN	-	71.560	47.190	54.707	41.656	47.291	38.989
	CR-CNN	-	72.016	**55.796**	**60.953**	43.961	58.893	45.060
	PCNN	-	**72.194**	50.791	57.687	44.667	54.703	44.011
	ATT-BLSTM	-	71.972	55.313	60.165	**45.336**	**60.004**	**45.928**
Bag2Sent	CNN+ONE	-	64.970	24.777	32.711	48.501	28.096	31.695
	CNN+ATT	-	**65.996**	22.729	30.976	50.334	26.239	30.488
	PCNN+ONE	-	64.020	**26.362**	**33.893**	**51.787**	**32.240**	**34.981**
	PCNN+ATT	-	63.542	24.388	31.913	48.728	28.334	32.367
Bag2Bag	CNN+ONE	0.671	**66.823**	**37.191**	**45.325**	43.478	45.078	39.539
	CNN+ATT	0.690	57.942	23.823	31.660	**50.632**	21.792	26.433
	PCNN+ONE	0.681	63.096	37.010	44.299	45.586	**47.206**	**41.843**
	PCNN+ATT	**0.699**	58.269	27.124	34.879	48.121	24.952	28.805

AUC & macro-F1 Scores among All Tracks

Conclusion

- We build the NYT-H dataset for accurate and consistent evaluation on distantly-supervised relation extraction task
- NYT-H can serve as a benchmark for Bag2Bag, Bag2Sent and Sent2Sent tracks
- We analyse the noise effects by distant supervision and offer a better way to evaluate the final models

Thanks

Q&A

Tong Zhu : tzhu7@stu.suda.edu.cn

https://github.com/Spico197/NYT-H