Efficient Argument Structure Extraction with Transfer Learning and Active Learning

Bloomberg

Engineering

Findings of ACL 2022

Xinyu Hua, AI Researcher, Bloomberg
Lu Wang, University of Michigan

TechAtBloomberg.com

Roadmap

- Motivation
- Prior Work
- Task and Model
- Dataset
- Transfer Learning
- Active Learning
- Conclusion

Motivation

Comment:

I think this submission does not meet the community standard.

The originality of the approach is unclear. Most existing work (…) The difference here is (…) not meaningful.

Secondly, none of the baselines uses (…), which is unfair comparison.

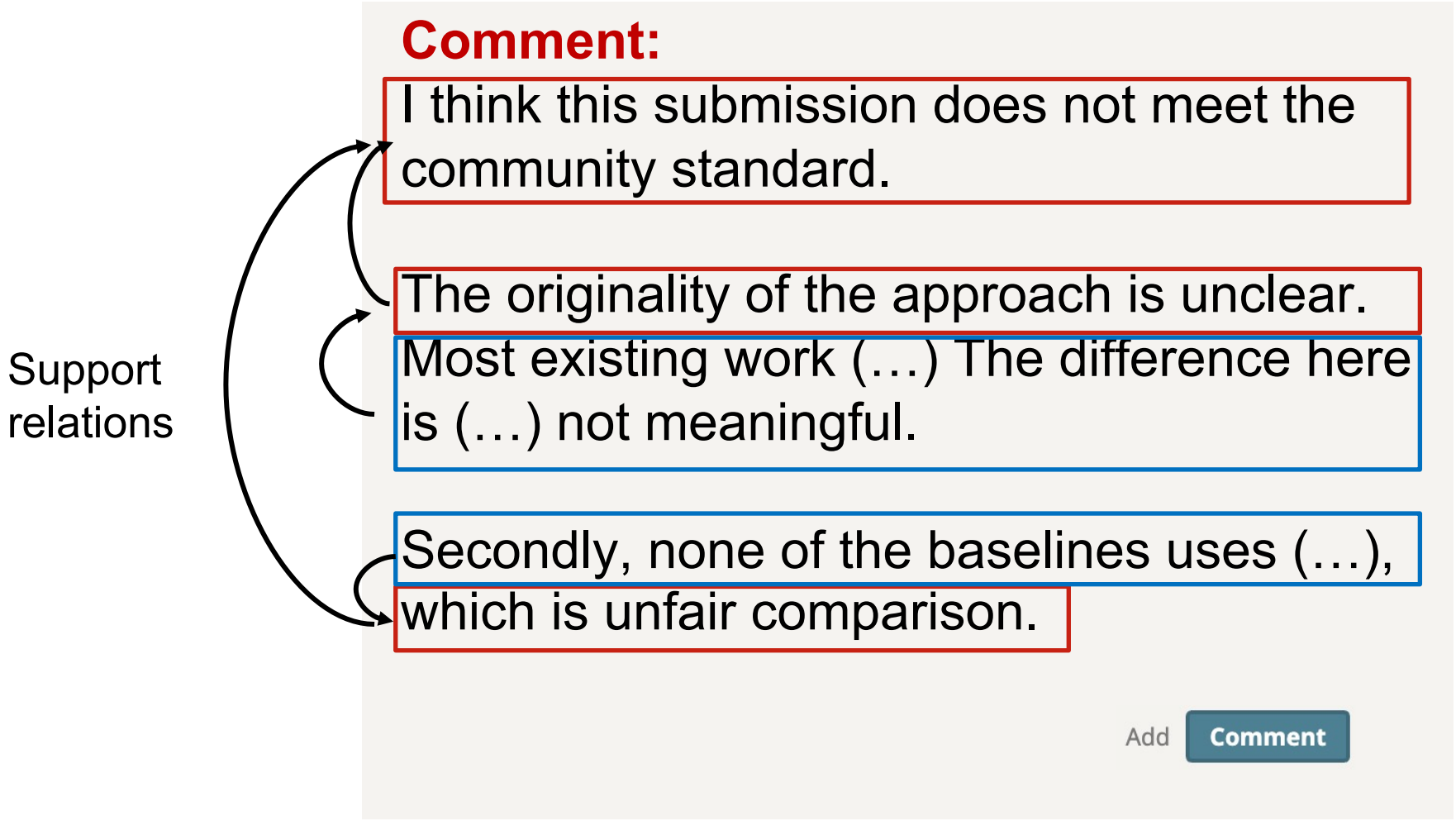
Add **Comment**

TechAtBloomberg.com

Bloomberg

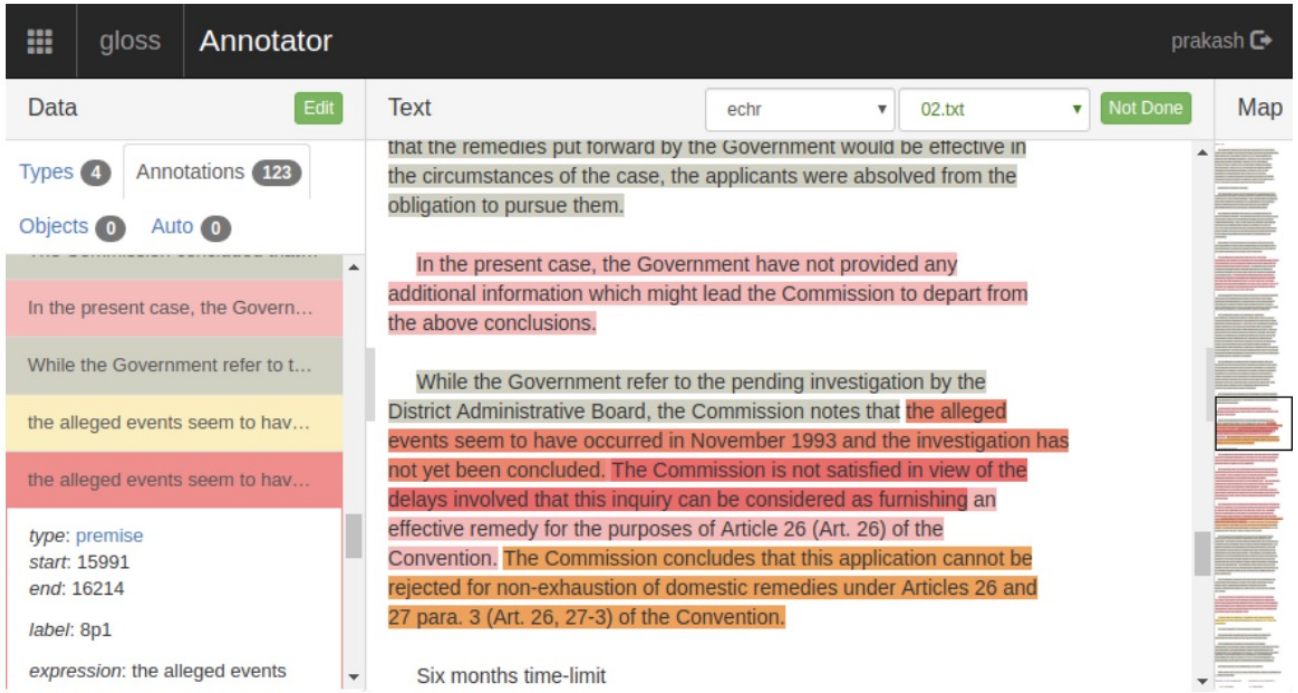
Engineering

Motivation



Motivation

- Annotating argument relations is difficult



Annotators (domain experts) need to scan through long documents.

Credit: Gloss interface, ECHR's Case Law [Poudyal+, 2020]

Motivation

- Many existing (small) datasets exist, but no unified framework

First, [cloning will be beneficial for many people who are in need of organ transplants]$_{Claim2}$. [Cloned organs will match perfectly to the blood group and tissue of patients]$_{Premise1}$ since [they can be raised from cloned stem cells of the patient]$_{Premise2}$. In addition, [it shortens the healing process]$_{Premise3}$. Usually, [it is very rare to find an appropriate organ donor]$_{Premise4}$ and [by using cloning in order to raise required organs the waiting time can be shortened tremendously]$_{Premise5}$.

Student essays [Stab & Gurevych, 2017]

(1) $400 is enough compensation,$_A$ as it can cover a one-way fare across the US.$_B$ I checked in a passenger on a $98.00 fare from east coast to Las Vegas the other day.$_C$

Online comments [Park & Cardie, 2018]

Example 2 *[True acupuncture was associated with 0.8 fewer hot flashes per day than sham at 6 weeks,]$_1$ [but the difference did not reach statistical significance (95% CI, -0.7 to 2.4;*

Biomedical domain, paper abstract [Mayer+, 2020]

"**The notion of security of person has not been given an independent interpretation** *(see in this respect Selçuk and Asker v. Turkey, nos. 23184/94 and 23185/94, Commission's report of 28 November 1996, §§ 185-187).*"

Legal domain, case law [Poudyal+, 2020]

TechAtBloomberg.com

Bloomberg

Engineering

Roadmap

- Motivation
- Prior Work
- Task and Model
- Dataset
- Transfer Learning
- Active Learning
- Conclusion

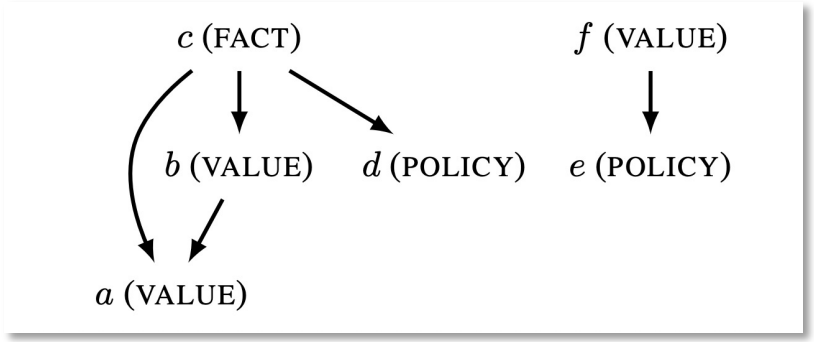
TechAtBloomberg.com

Bloomberg

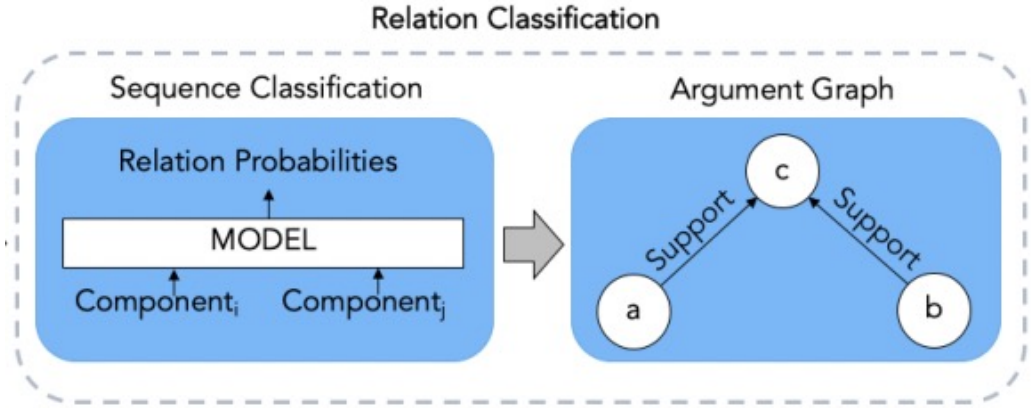
Engineering

Prior Work

- Argument Structure Prediction



Factor graph with constraints
[Niculae, Park, and Cardie, 2017]



Pairwise predictions
[Stab and Gurevych, 2017]
[Mayer, Cabrio, and Villata, 2020]

Prior Work

- Transfer Learning for Structured Prediction
  - Transductive
  - Inductive

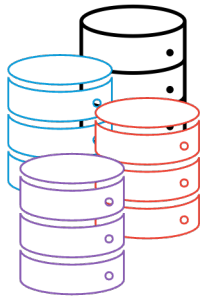
Domain: source
Data: labeled

Domain: target
Data: labeled

Both source and target data are labeled (supervised training), on the **same task**

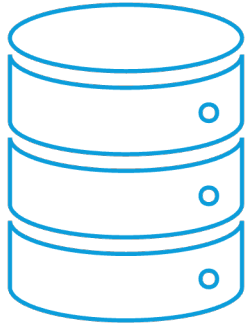
Prior Work

- Transfer Learning for Structured Prediction
  - Transductive
  - Inductive



Domain: open
Data: unlabeled

OR



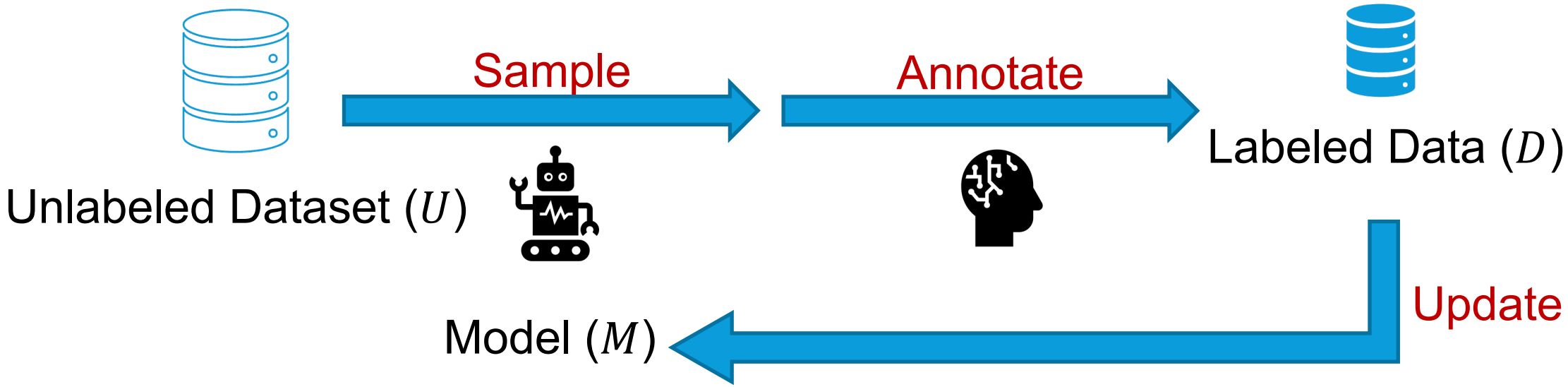
Domain: target
Data: unlabeled

Domain: target
Data: labeled

Transfer from unlabeled data, e.g., using self-supervised training

Prior Work

- Active Learning [Settles, 2009; Aggarwal+, 2014]
  - Unlabeled dataset is available
  - Annotation is subject to a budget
  - Goal is to select the most informative samples



Roadmap

- Motivation
- Prior Work
- Task and Model
- Dataset
- Transfer Learning
- Active Learning
- Conclusion

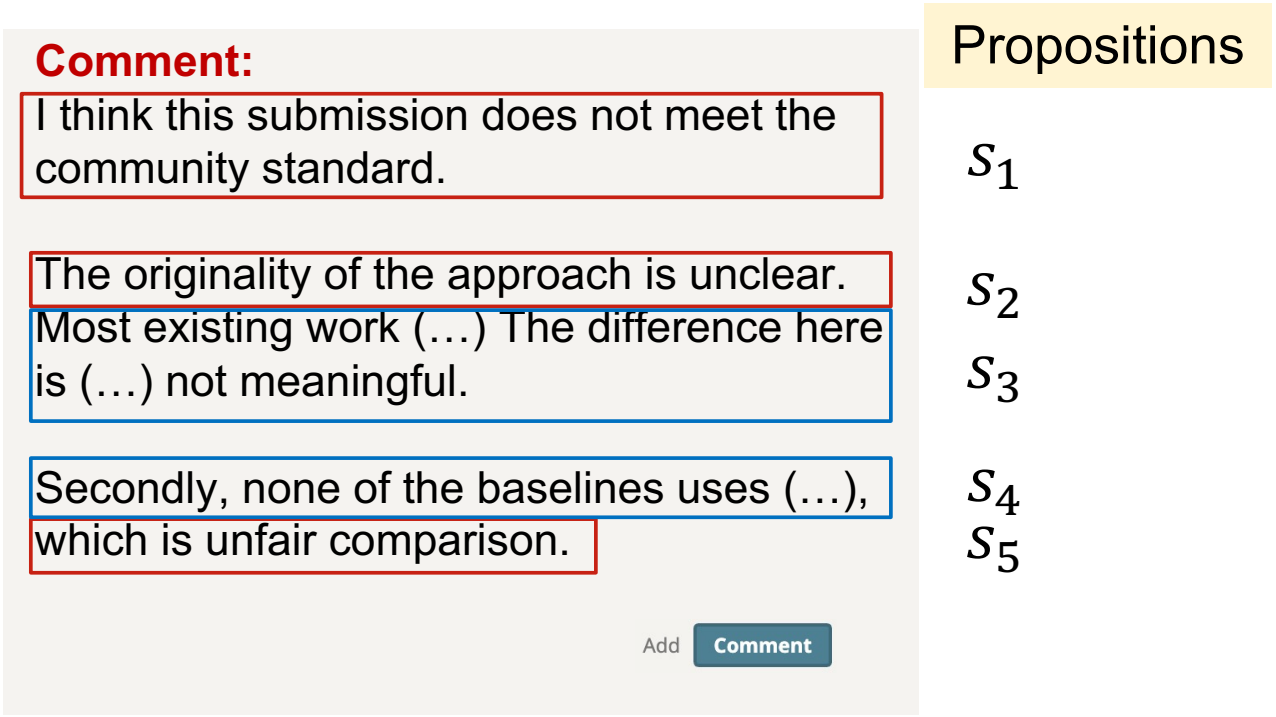
TechAtBloomberg.com

Bloomberg

Engineering

Task and Model

- Task Formulation



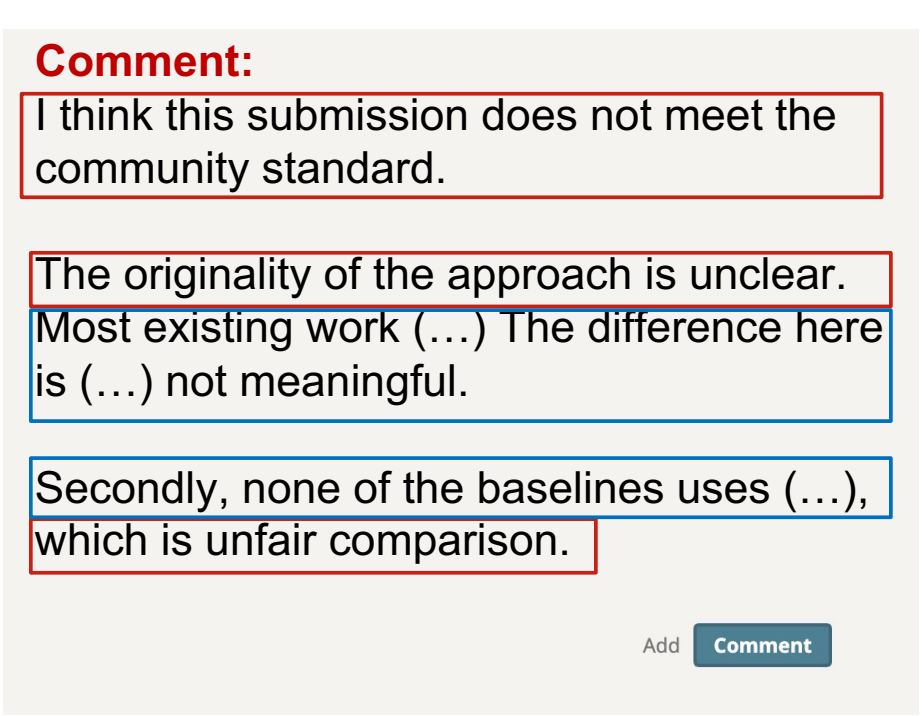
Task and Model

- Task Formulation

To (head)

From (tail)

	s_1	s_2	s_3	s_4	s_5
s_1					
s_2	✓				
s_3		✓			
s_4					✓
s_5	✓				



Propositions

s_1

s_2

s_3

s_4

s_5

TechAtBloomberg.com

Bloomberg

Engineering

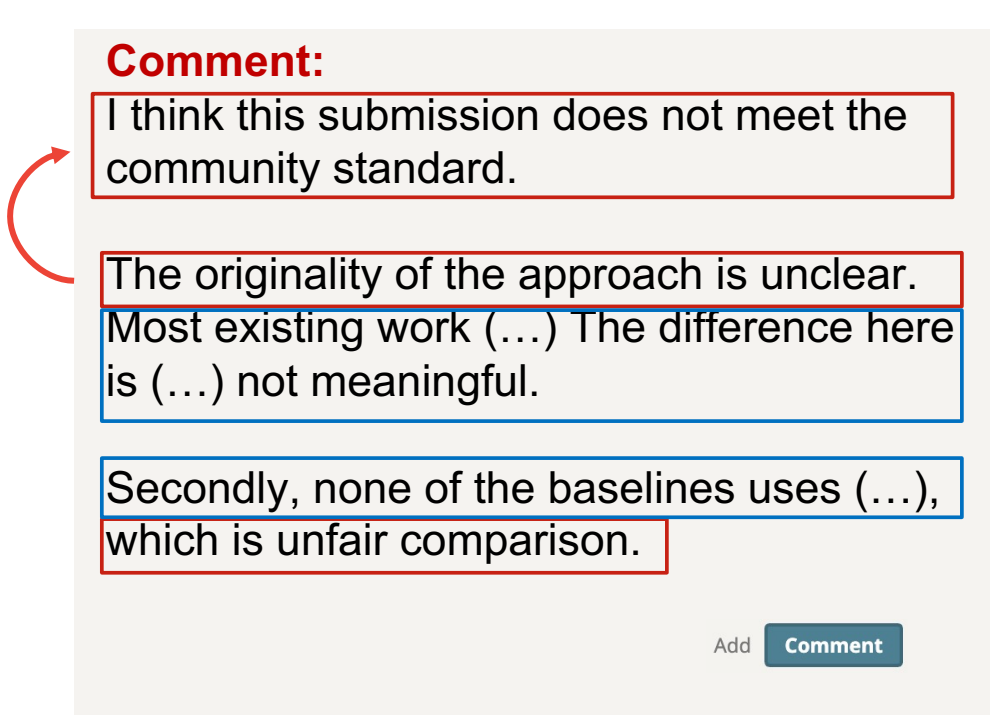
Task and Model

- Task Formulation

To (head)

From (tail)

	s_1	s_2	s_3	s_4	s_5
s_1					
s_2	✓				
s_3		✓			
s_4					✓
s_5	✓				



TechAtBloomberg.com

Bloomberg

Engineering

Task and Model

- Task Formulation

To (head)

From (tail)

	s_1	s_2	s_3	s_4	s_5
s_1					
s_2	✓				
s_3		✓			
s_4					✓
s_5	✓				

Comment:

I think this submission does not meet the community standard.

The originality of the approach is unclear.

Most existing work (…) The difference here is (…) not meaningful.

Secondly, none of the baselines uses (…), which is unfair comparison.

Add **Comment**

Propositions

s_1

s_2

s_3

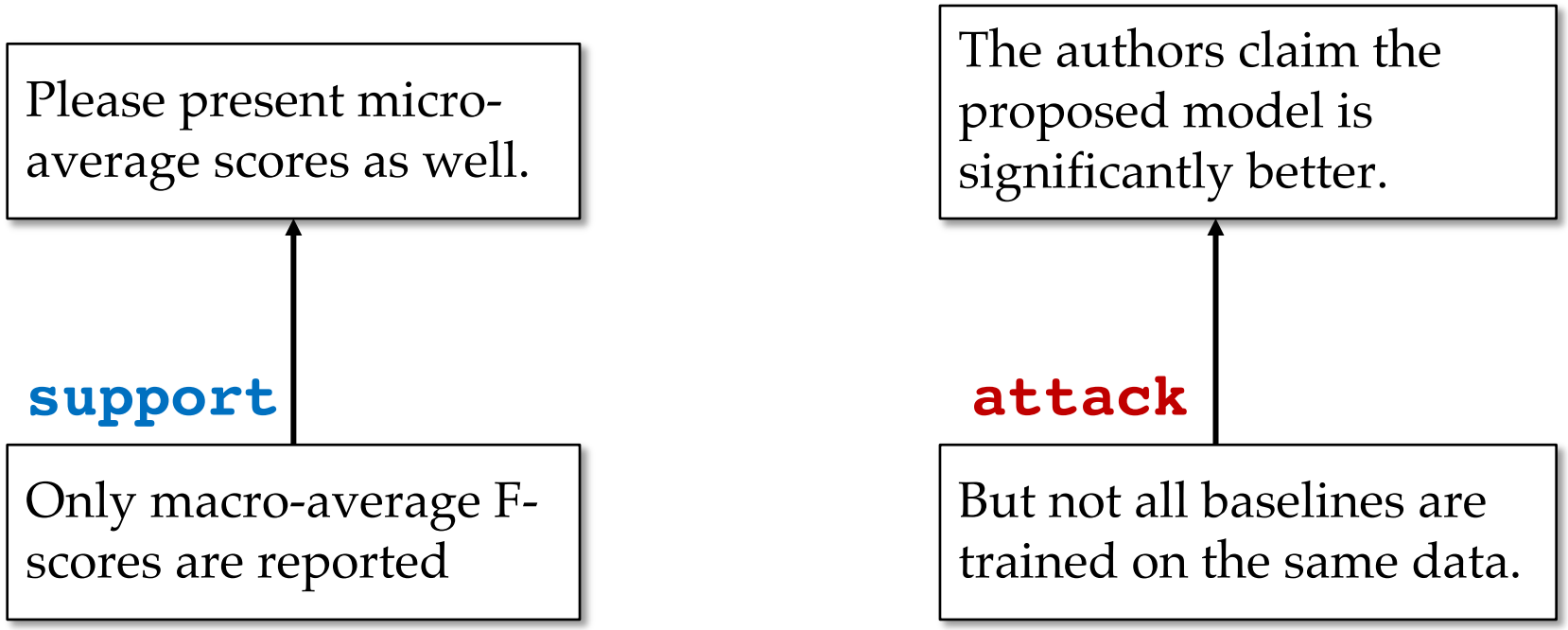
s_4

s_5

Simplified setting assumes head propositions (s_1, s_2, s_5) are given.

Task and Model

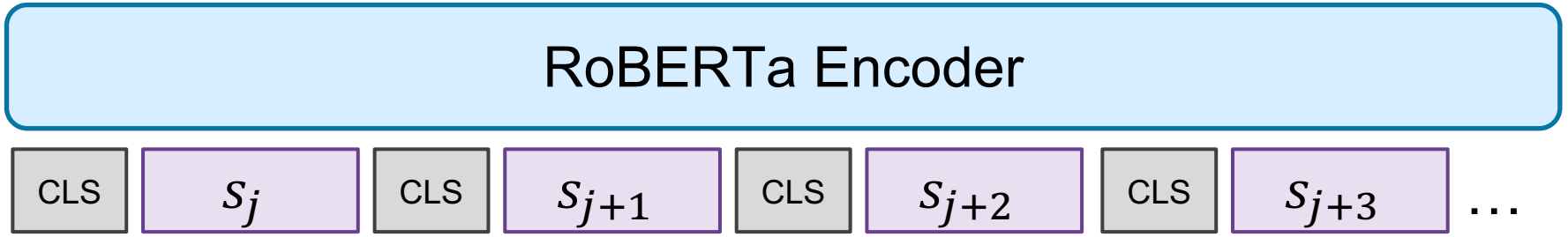
- Task Formulation
  - Types of relations: `support` and `attack`



Task and Model

- Context-aware model

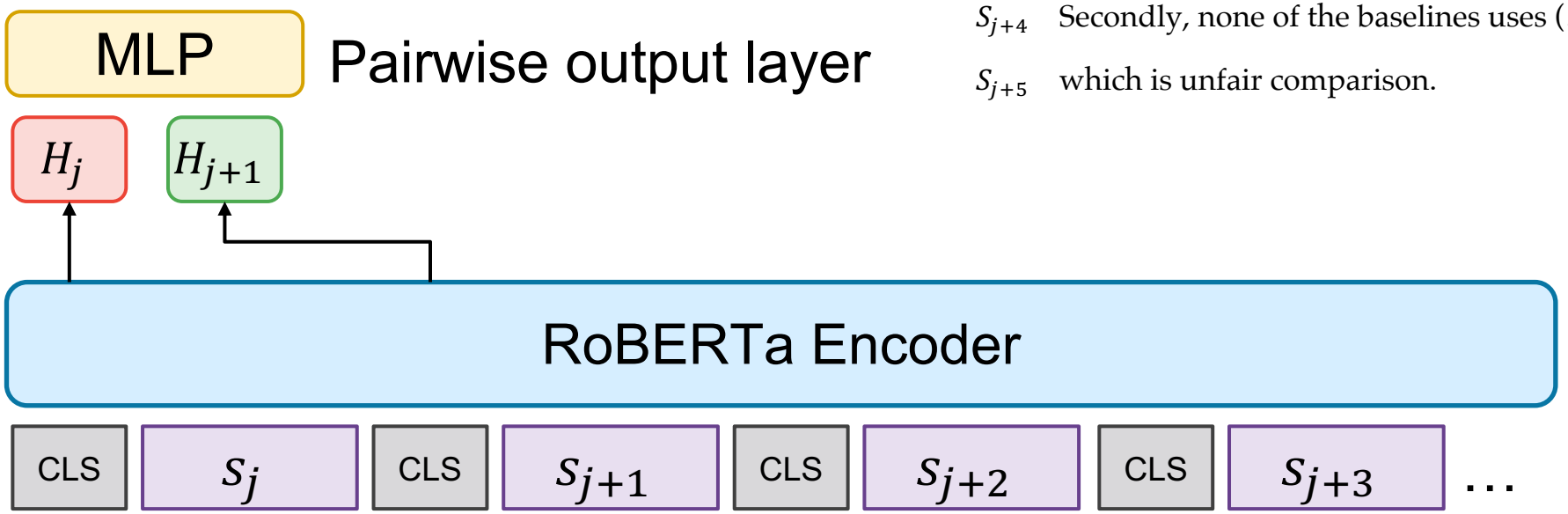
S_j	I think this submission does not meet the community standard.
S_{j+1}	The originality of the approach is unclear.
S_{j+2}	Most existing work (…)
S_{j+3}	The difference here is (…) not meaningful.
S_{j+4}	Secondly, none of the baselines uses (…),
S_{j+5}	which is unfair comparison.



Task and Model

- Context-aware model

$P(S_j \leftarrow S_{j+1})$

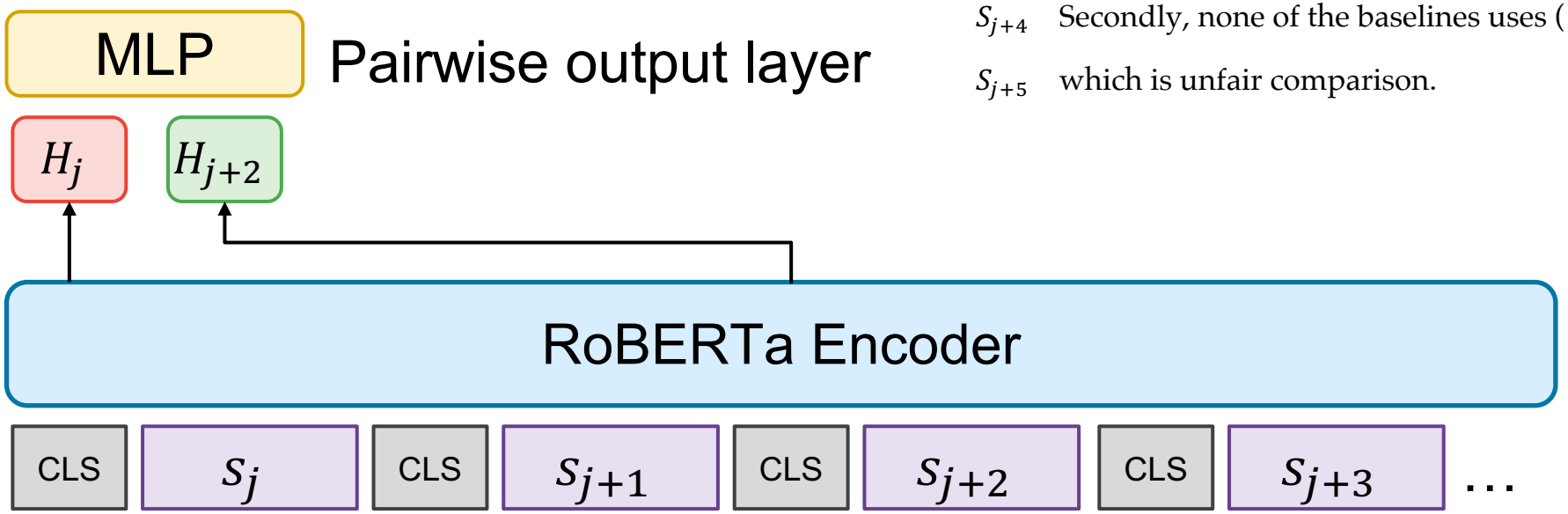


S_j	I think this submission does not meet the community standard.
S_{j+1}	The originality of the approach is unclear.
S_{j+2}	Most existing work (…)
S_{j+3}	The difference here is (…) not meaningful.
S_{j+4}	Secondly, none of the baselines uses (…),
S_{j+5}	which is unfair comparison.

Task and Model

- Context-aware model

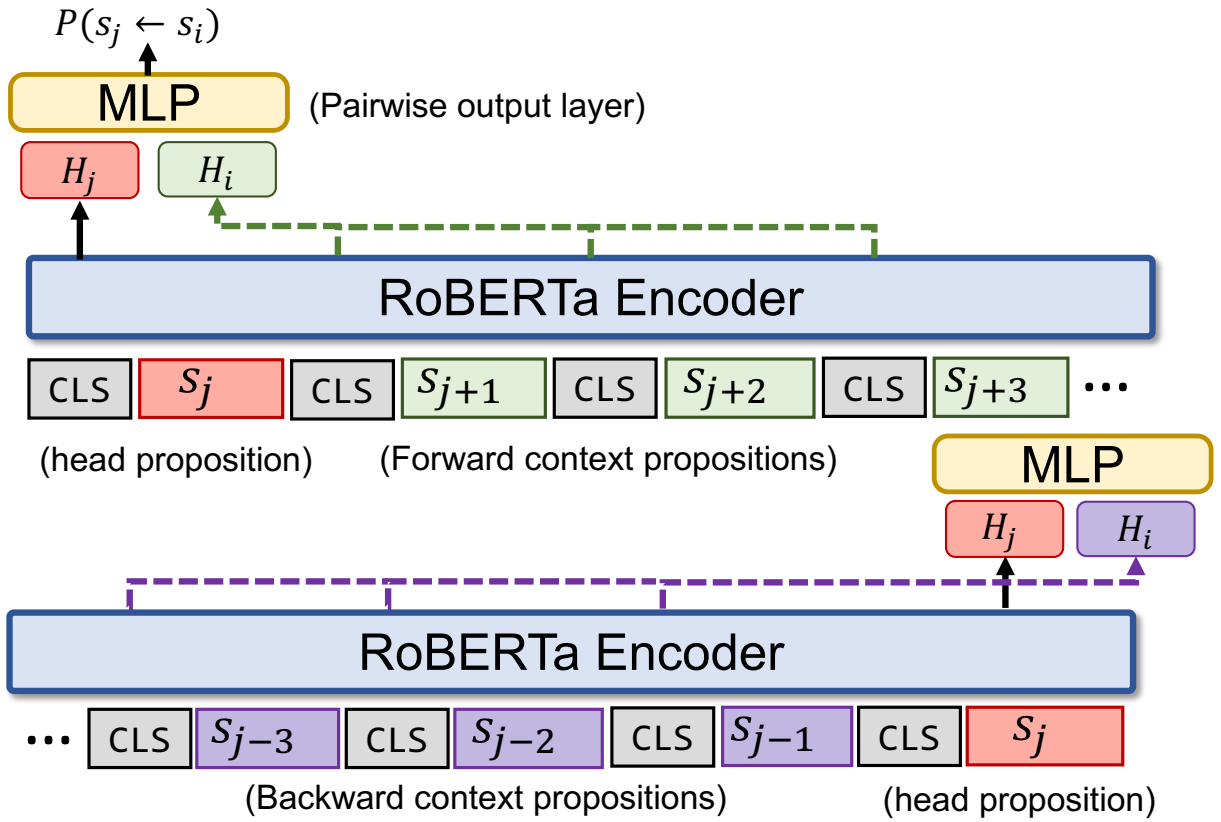
$P(S_j \leftarrow S_{j+2})$



S_j	I think this submission does not meet the community standard.
S_{j+1}	The originality of the approach is unclear.
S_{j+2}	Most existing work (…)
S_{j+3}	The difference here is (…) not meaningful.
S_{j+4}	Secondly, none of the baselines uses (…),
S_{j+5}	which is unfair comparison.

Task and Model

- Context-aware model



Roadmap

- Motivation
- Prior Work
- Task and Model
- Dataset
- Transfer Learning
- Active Learning
- Conclusion

TechAtBloomberg.com

Bloomberg

Engineering

Dataset

- AMPERE++ (new annotation)
  - Domain: paper reviews from openreview.net
  - Originally collected in our prior work [Hua+, 2019]
  - 3,636 argument relations (`support` and `attack`)
  - IAA: 0.654 (Fleiss' kappa)

Dataset

- Essays [Stab & Gurevych, 2017]
- AbstRCT [Mayer+, 2020]
- ECHR [Poudyal+, 2020]
- CDCP [Park & Cardie, 2018]

TechAtBloomberg.com

Bloomberg

Engineering

Dataset

- Essays [Stab & Gurevych, 2017]

First, [cloning will be beneficial for many people who are in need of organ transplants]$_{Claim2}$. [Cloned organs will match perfectly to the blood group and tissue of patients]$_{Premise1}$ since [they can be raised from cloned stem cells of the patient]$_{Premise2}$. In addition, [it shortens the healing process]$_{Premise3}$. Usually, [it is very rare to find an appropriate organ donor]$_{Premise4}$ and [by using cloning in order to raise required organs the waiting time can be shortened tremendously]$_{Premise5}$.

- AbstRCT [Mayer+, 2020]
- ECHR [Poudyal+, 2020]
- CDCP [Park & Cardie, 2018]

TechAtBloomberg.com

Bloomberg

Engineering

Dataset

- Essays [Stab & Gurevych, 2017]
- AbstRCT [Mayer+, 2020]
- ECHR [Poudyal+, 2020]
- CDCP [Park & Cardie, 2018]

Example 2 *[True acupuncture was associated with 0.8 fewer hot flashes per day than sham at 6 weeks,]$_1$ [but the difference did not reach statistical significance (95% CI, -0.7 to 2.4; P = .3).]$_2$*

Dataset

- Essays [Stab & Gurevych, 2017]
- AbstRCT [Mayer+, 2020]
- ECHR [Poudyal+, 2020]
- CDCP [Park & Cardie, 2018]

> **"The notion of security of person has not been given an independent interpretation** *(see in this respect Selçuk and Asker v. Turkey, nos. 23184/94 and 23185/94, Commission's report of 28 November 1996, §§ 185-187)."*

TechAtBloomberg.com

Bloomberg

Engineering

Dataset

- Essays [Stab & Gurevych, 2017]
- AbstRCT [Mayer+, 2020]
- ECHR [Poudyal+, 2020]
- CDCP [Park & Cardie, 2018]

(1) $400 is enough compensation,$_A$ as it can cover a one-way fare across the US.$_B$ I checked in a passenger on a $98.00 fare from east coast to Las Vegas the other day.$_C$

TechAtBloomberg.com

Bloomberg

Engineering

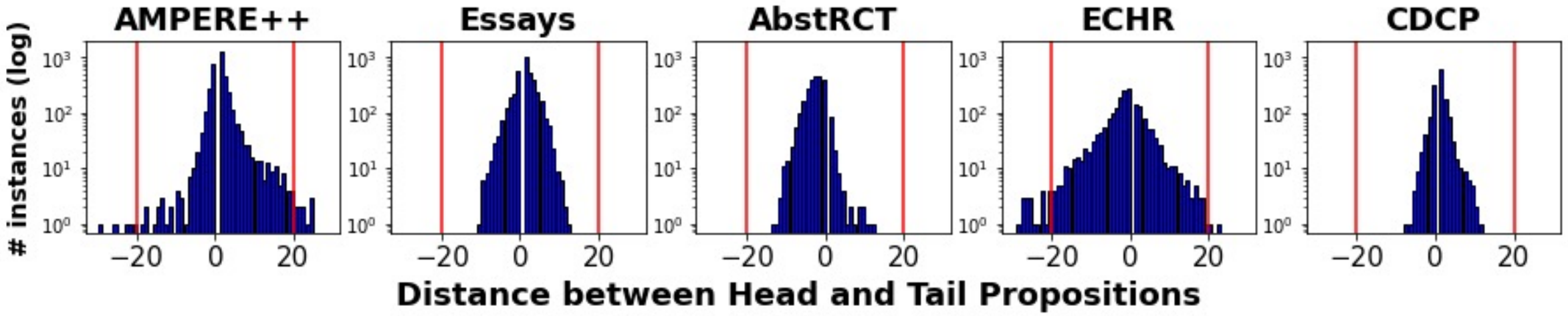
Dataset

- Statistics

	AMPERE++	Essays	AbstRCT	ECHR	CDCP
# Documents	400	402	700	42	731
# Propositions	10.4K	12.4K	5.7K	6.3K	4.9K
# Support Rel.	3,370	3,613	2,402	1,946	1,426
# Attack Rel.	266	219	70	0	0
# Head Prop.	2,268	1,707	1,138	741	1,037
Density	22%	14%	20%	12%	21%

Dataset

- Statistics: distribution of head-tail distance



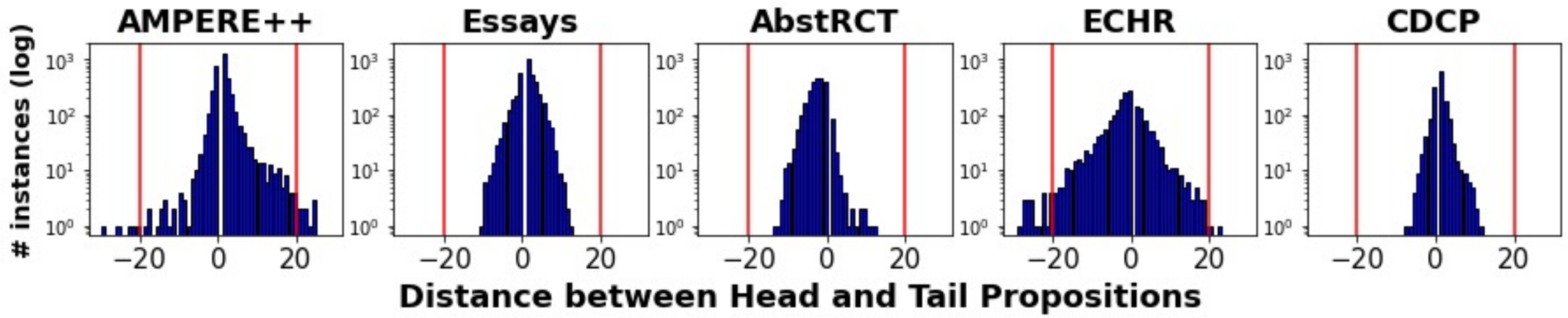
TechAtBloomberg.com

Bloomberg

Engineering

Dataset

- Statistics: distribution of head-tail distance



Most relations span less than 20 propositions.

TechAtBloomberg.com

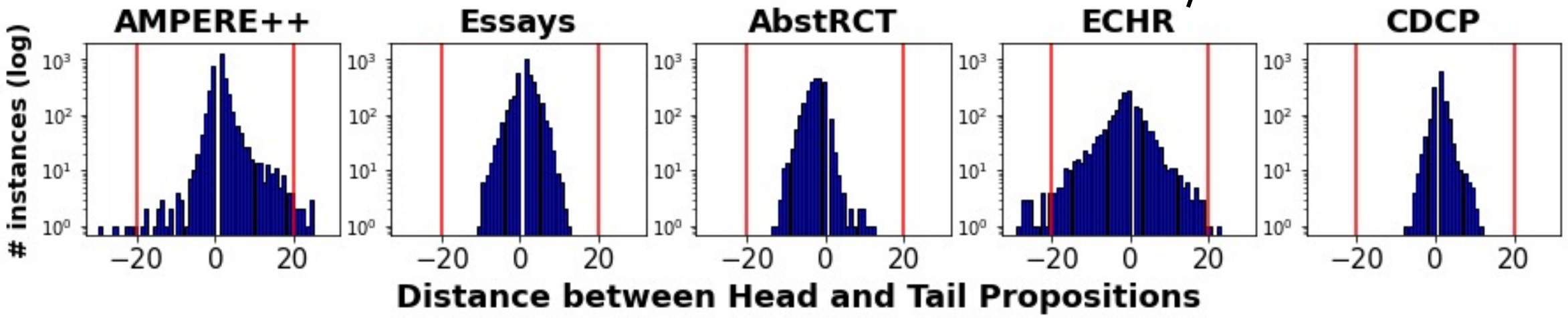
Bloomberg

Engineering

Dataset

- Statistics: distribution of head-tail distance

Long-term relations are more common in the legal domain

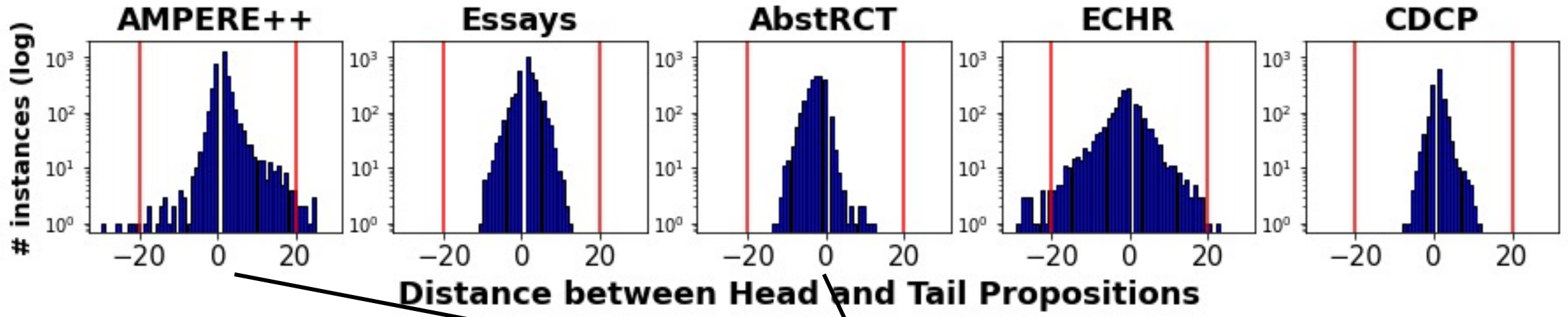


Bloomberg

Engineering

Dataset

- Statistics: distribution of head-tail distance



Review: verdicts first, support later
Paper abstract: evidence first, conclusion later

Bloomberg

Engineering

Standard Supervised Setting

- Baselines and our context-aware model
- Macro F1 scores

	AMPERE++	Essays	AbstRCT	ECHR	CDCP
SVM-linear	24.82	28.69	33.60	21.18	29.01
SVM-RBF	26.38	31.68	32.65	21.36	30.34
SEQPAIR	23.40	38.37	**66.96**	13.76	35.23
OURS (head given)	**77.64**	**71.30**	63.62	**70.82**	**70.37**
OURS (end-to-end)	74.34	67.68	63.73	61.35	63.13

TechAtBloomberg.com

Bloomberg

Engineering

Standard Supervised Setting

- Baselines and our context-aware model
- Macro F1 scores

	AMPERE++	Essays	AbstRCT	ECHR	CDCP
SVM-linear	24.82	28.69	33.60	21.18	29.01
SVM-RBF	26.38	31.68	32.65	21.36	30.34
SEQPAIR	23.40	38.37	**66.96**	13.76	35.23
OURS (head given)	**77.64**	**71.30**	63.62	**70.82**	**70.37**
OURS (end-to-end)	74.34	67.68	63.73	61.35	63.13

Takeaways:
1) Context-aware model is generally much better
2) End-to-end and simplified (head given) setting are close

Roadmap

- Motivation
- Prior Work
- Task and Model
- Dataset
- Transfer Learning
- Active Learning
- Conclusion

TechAtBloomberg.com

Bloomberg

Engineering

Transfer Learning

- Transductive TL
  - Same task, different domains (datasets)
  - (Source) model weights as (target) initialization

Transfer Learning

- Transductive TL
  - Same task, different domains (datasets)
  - (Source) model weights as (target) initialization

Target Domain

Source Domain	AMPERE++	Essays	AbstRCT	ECHR	CDCP
AMPERE++		73.84	63.42	76.50	75.93
Essays	77.93		60.62	68.72	74.11
AbstRCT	76.29	71.17		73.31	69.17
ECHR	77.69	70.82	47.91		69.30
CDCP	77.87	68.37	62.38	72.03	

Transfer Learning

- Transductive TL
  - Same task, different domains (datasets)
  - (Source) model weights as (target) initialization

Transfer settings that outperform standard supervised setting are **highlighted**

Target Domain

Source Domain	AMPERE++	Essays	AbstRCT	ECHR	CDCP
AMPERE++		**73.84**	63.42	**76.50**	**75.93**
Essays	**77.93**		60.62	68.72	**74.11**
AbstRCT	76.29	71.17		**73.31**	69.17
ECHR	**77.69**	70.82	47.91		69.30
CDCP	**77.87**	68.37	62.38	**72.03**	

Roadmap

- Motivation
- Prior Work
- Task and Model
- Dataset
- Transfer Learning
- Active Learning
- Conclusion

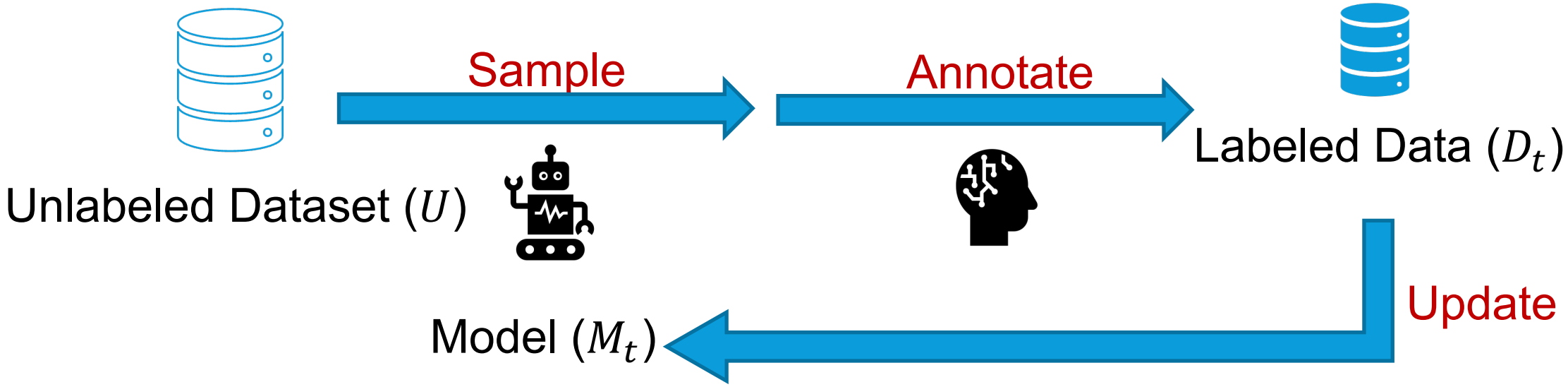
TechAtBloomberg.com

Bloomberg

Engineering

Active Learning

- Experiment settings:
  - 10 iterations, 500 samples per iteration



Active Learning

- Acquisition strategies
  - RANDOM
  - MAX-ENTROPY [Lewis & Gale, 1994; Joshi+, 2009]
  - BALD [Houlsby+, 2011]
  - CORESET [Sener & Savarese, 2018]

Active Learning

- Acquisition strategies
  - RANDOM
  - MAX-ENTROPY [Lewis & Gale, 1994; Joshi+, 2009]
  - BALD [Houlsby+, 2011]

  } Picks the most uncertain samples

  - CORESET [Sener & Savarese, 2018] → Maximizes sample diversity

Active Learning

- Acquisition strategies
  - RANDOM
  - MAX-ENTROPY [Lewis & Gale, 1994; Joshi+, 2009]
  - BALD [Houlsby+, 2011]
  - CORESET [Sener & Savarese, 2018]
  - NOVEL-VOCAB

$$\text{novelty-score}(\boldsymbol{s}_i) = \sum_{w_t \in \boldsymbol{s}_i} \frac{f_{i,t}}{(1 + \mathcal{V}(w_t))}$$

Frequency of word w_t in sample S_i

Frequency of word w_t in the labeled D_t

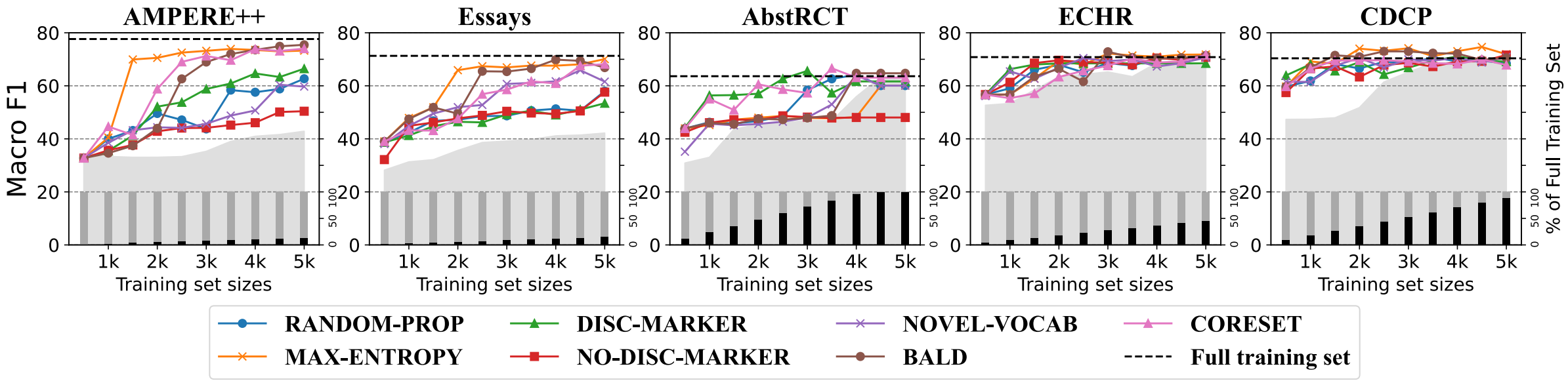
Active Learning

- Acquisition strategies
  - RANDOM
  - MAX-ENTROPY [Lewis & Gale, 1994; Joshi+, 2009]
  - BALD [Houlsby+, 2011]
  - CORESET [Sener & Savarese, 2018]
  - NOVEL-VOCAB
  - DISC-MARKER

because	therefore	however
although	though	nevertheless
nonetheless	thus	hence
consequently	for this reason	due to
in particular	particularly	specifically
in fact	actually	but

Active Learning

- Results



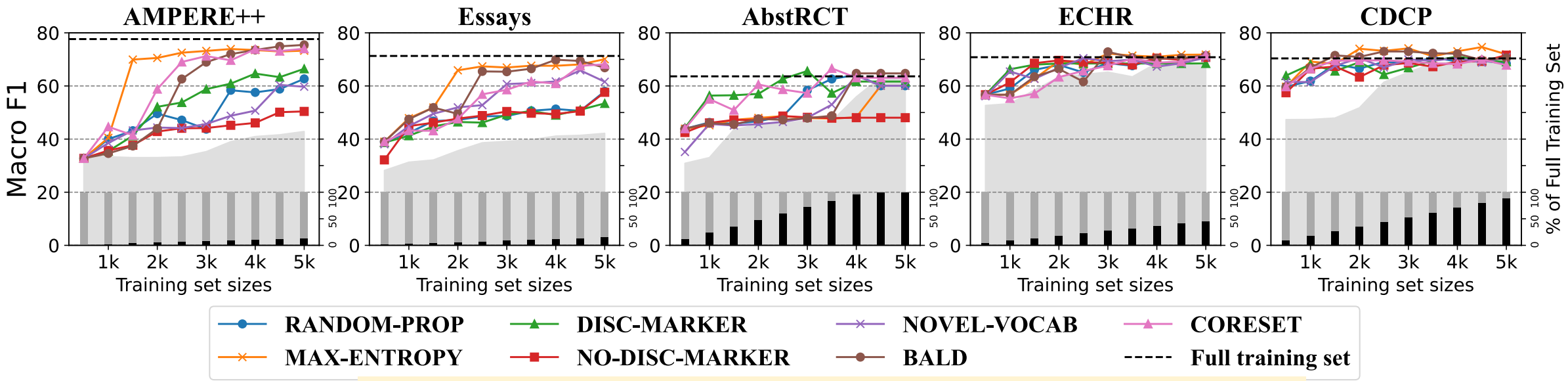
TechAtBloomberg.com

Bloomberg

Engineering

Active Learning

- Results



Takeaways:

1) AL helps more in the early stage (low-resource).
2) Model-independent strategies achieve competitive performance, yet they are much faster.

TechAtBloomberg.com

Bloomberg

Engineering

Conclusion

- We present a simple yet effective framework for argument structure extraction.
- We release AMPERE++, a newly annotated dataset on peer reviews.
- We showcase two data efficient learning methods (transfer learning and active learning) using our model.

Questions?


https://xinyuhua.github.io/Resources/acl22/


https://zenodo.org/record/6362430


xhua22@bloomberg.net

Bloomberg

Engineering

TechAtBloomberg.com

Transfer Learning

- Inductive TL
  - Same domain, different tasks (self-supervision)
  - MLM: masked language model
  - Context-Pert: context-aware sentence perturbation

	AMPERE++	Essays	AbstRCT
MLM	**78.10**	**74.21**	**64.48**
Context-Pert	**79.01**	68.36	59.47

Active Learning

- Acquisition strategies
  - RANDOM: RANDOM-CTX vs. RANDOM-PROP
  - MAX-ENTROPY [Lewis & Gale, 1994; Joshi+, 2009]
  - BALD [Houlsby+, 2011]
  - CORESET [Sener & Savarese, 2018]