Northeastern University
Khoury College of
Computer Sciences

Argument Mining for Understanding Peer Reviews

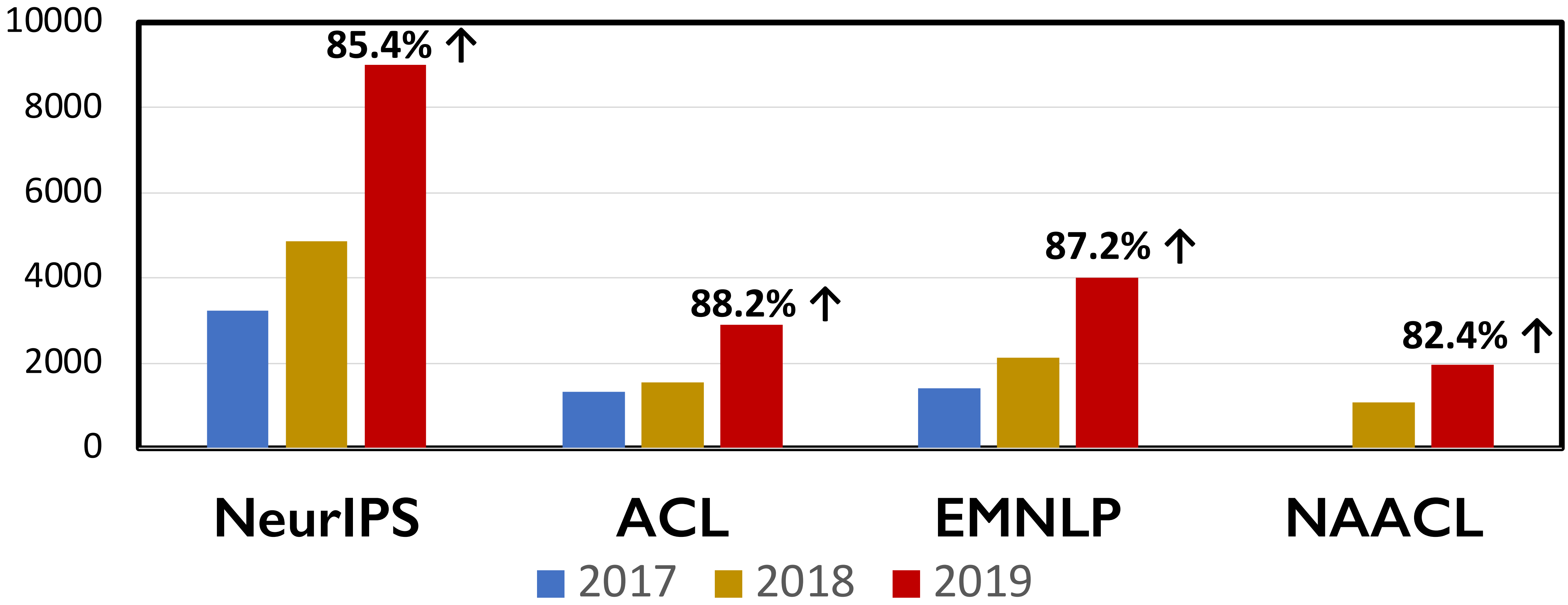
Xinyu Hua, Mitko Nikolov, Nikhil Badugu, Lu Wang

Project page: https://xinyuhua.github.io/Resources/naacl19

June 4, 2019
NAACL
Minneapolis

Some recent developments in ML/NLP conference...

Submissions



Source: https://aclweb.org/aclwiki/Conference_acceptance_rates

Challenges

- Huge efforts required for the reviewers

Motivation

- Growing interests in understanding peer-reviews

Motivation

- Growing interests in understanding peer-reviews
  - Assess rebuttal and author response [Gao, Eger, Kuznetsov, Gurevych, and Miyao, 2019]
  - Distinguish high/low quality review [Falkenberg and Soranno, 2018]
  - Predicting acceptance from review [Kang, Ammar, Dalvi, van Zuylen, Kohlmeier, Hovy, and Schwartz, 2018]

Motivation

- Reviews resemble arguments

Rating: 6: Marginally above acceptance threshold

Review: *This paper proposes to bring together multiple inductive biases... The human evaluation is straight-forward and meaningful... I would like this point to be clarified better in the paper. I think showing results on grounded generation tasks like...would make a stronger case...*

URL: https://openreview.net/revisions?id=HkN9lyRxG

Motivation

- Reviews resemble arguments

Summary of the paper

Rating: 6: Marginally above acceptance threshold

Review: *This paper proposes to bring together multiple inductive biases... The human evaluation is straight-forward and meaningful... I would like this point to be clarified better in the paper. I think showing results on grounded generation tasks like...would make a stronger case...*

URL: https://openreview.net/revisions?id=HkN9lyRxG

Motivation

- Reviews resemble arguments

Subjective judgement

Rating: 6: Marginally above acceptance threshold

Review: *This paper proposes to bring together multiple inductive biases... The human evaluation is straight-forward and meaningful... I would like this point to be clarified better in the paper. I think showing results on grounded generation tasks like...would make a stronger case...*

URL: https://openreview.net/revisions?id=HkN9lyRxG

Motivation

- Reviews resemble arguments

Suggestions

Rating: 6: Marginally above acceptance threshold

Review: *This paper proposes to bring together multiple inductive biases... The human evaluation is straight-forward and meaningful... I would like this point to be clarified better in the paper. I think showing results on grounded generation tasks like...would make a stronger case...*

URL: https://openreview.net/revisions?id=HkN9lyRxG

Motivation

- Prior work on argument mining
  - Claim/Premise detection [Persing and Ng, 2016; Stab and Gurevych, 2017; Shnarch et al, 2018]

Motivation

- Prior work on argument mining
  - Claim/Premise detection [Persing and Ng, 2016; Stab and Gurevych, 2017; Shnarch et al, 2018]

> *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]$_{Premise5}$.*

Stab and Gurevych (2017)

Motivation

- Prior work on argument mining
  - Claim/Premise detection [Persing and Ng, 2016; Stab and Gurevych, 2017; Shnarch et al, 2018]
  - Argument classification [Niculae, Park, and Cardie, 2017; Habernal and Gurevych, 2017; Hidey, Musi, Hwang, Muresan, and McKeown, 2017]

Motivation

Toulmin model
Credit: Habernal and Gurevych (2017)

Claim is an assertion put forward publicly for general acceptance (Toulmin, Rieke, and Janik 1984, page 29) or the conclusion we seek to establish by our arguments (Freeley and Steinberg 2008, page 153).

Data (Grounds) This is the evidence to establish the foundation of the claim (Schiappa and Nordin 2013) or, as simply put by Toulmin, "the data represent what we have to go on" (Toulmin 2003, page 90). The name of this concept was later changed to *grounds* in Toulmin, Rieke, and Janik (1984).

Warrant The role of *warrant* is to justify a logical inference from the *grounds* to the *claim*.

Backing is a set of information that stands behind the *warrant*. It assures its trustworthiness.

Qualifier limits the degree of certainty under which the argument should be accepted. It is the degree of force that the *grounds* confer on the *claim* in virtue of the *warrant* (Toulmin 2003, page 93).

Rebuttal presents a situation in which the *claim* might be defeated.

Motivation

- Prior work on argument mining
  - Claim/Premise detection [Persing and Ng, 2016; Stab and Gurevych, 2017; Shnarch et al, 2018]
  - Argument classification [Niculae, Park, and Cardie, 2017; Habernal and Gurevych, 2017; Hidey, Musi, Hwang, Muresan, and McKeown, 2017]
  - Argument structures [Stab and Gurevych, 2016; Persing and Ng, 2016; Niculae, Park, and Cardie, 2017]

Motivation

- Our goal: Apply existing argument mining tools to understand peer-review quality

Motivation

- Our goal: Apply existing argument mining tools to understand peer-review quality
- Initial step: argument component analysis

Motivation

- Our goal: Apply existing argument mining tools to understand peer-review quality
- Initial step: argument component analysis
- Future work: review structure analysis

Roadmap

- Motivation
- Argument Components
- Annotation
- Experiment
- Analysis
- Conclusion

Roadmap

- Motivation
- Argument Components
- Annotation
- Experiment
- Analysis
- Conclusion

Argument Components

- Goal: To classify arguments by their functions and subjectivity

Argument Components

- Goal: To classify arguments by their functions and subjectivity

  - Evaluation: subjective judgements
  - Request: suggestions
  - Fact: objective and verifiable
  - Reference: citations and URLs
  - Quote: direct quotation from the paper

"This paper is novel and interesting"

Argument Components

- Goal: To classify arguments by their functions and subjectivity

  - Evaluation: subjective judgements
  - Request: suggestions
  - Fact: objective and verifiable
  - Reference: citations and URLs
  - Quote: direct quotation from the paper

"More baselines should be added"

Argument Components

- Goal: To classify arguments by their functions and subjectivity

  - Evaluation: subjective judgements
  - Request: suggestions
  - Fact: objective and verifiable
  - Reference: citations and URLs
  - Quote: direct quotation from the paper

"The authors propose an attention based method."

Argument Components

- Goal: To classify arguments by their functions and subjectivity

- Evaluation: subjective judgement
- Request: suggestions
- Fact: objective and verifiable
- Reference: citations and URLs
- Quote: direct quotation from the paper

MidiNet (Yang et al); “In sec 2: ‘we experiment with …’”

Roadmap

- Motivation
- Argument Components
- Annotation
- Experiment
- Analysis
- Conclusion

AMPERE: Argument Mining for PEer REviews

- Data: 400 reviews randomly sampled from ICLR 2018
- Average # words: 477.3
- Average # sentences: 20.1

Open
Review
.net

AMPERE: annotation

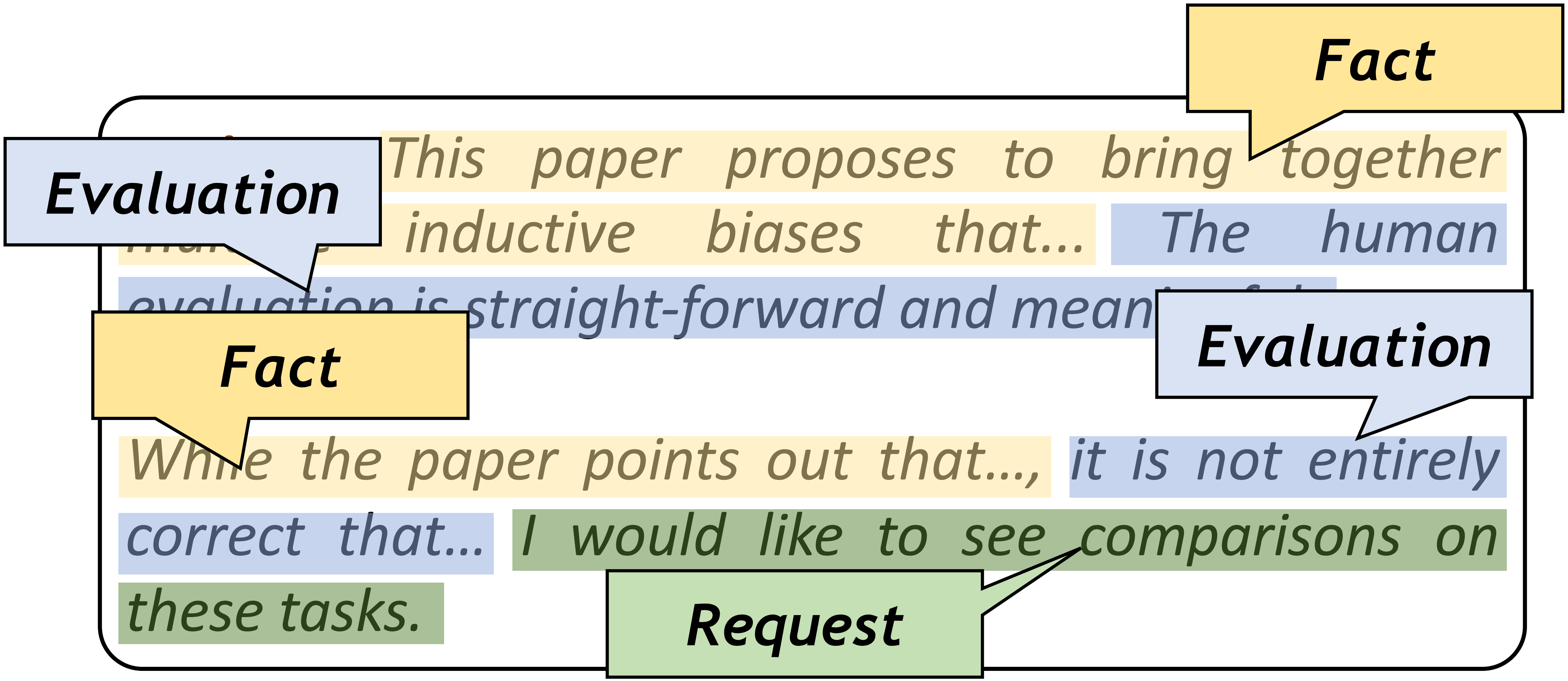
- Task I: proposition segmentation
- Task II: proposition classification

Review: *This paper proposes to bring together multiple inductive biases that... The human evaluation is straight-forward and meaningful...*

While the paper points out that..., it is not entirely correct that... I would like to see comparisons on these tasks.

Review: *This paper proposes to bring together multiple inductive biases that... The human evaluation is straight-forward and meaningful...*

While the paper points out that..., it is not entirely correct that... I would like to see comparisons on these tasks.


Fact
Evaluation
This paper proposes to bring together
inductive biases that...
The human
evaluation is straight-forward and mean
Fact
Evaluation
While the paper points out that...,
it is not entirely
correct that...
I would like to see comparisons on
these tasks.
Request

AMPERE: annotation

- Statistics

Krippendorff's α: 0.61
Cohen's κ: 0.64

Evaluation	Request	Fact	Reference	Quote	Non-Arg	Total
3,982	1,911	3,786	207	161	339	10,386

Roadmap

- Motivation
- Argument Components
- Annotation
- Experiment
- Analysis
- Conclusion

Experiment

- Data split:
  - Training: 320 reviews (7,999 propositions)
  - Test: 80 reviews (2,387 propositions)
  - Hyper-parameter tuning: 5-fold cross validation on training set

Experiment

- Data split
- Task I: segmentation (BIO tagging)
  - Model 1: CRF with features from Stab and Gurevych (2017)
  - Model 2: BiLSTM-CRF with ELMo [Huang, Xu, and Yu, 2015; Ma and Hovy, 2016; Peters et al, 2018]

Experiment

- Data split
- Task I: segmentation (BIO tagging)
  - Model 1: CRF with features from Stab and Gurevych (2017)
  - Model 2: BiLSTM-CRF with ELMo [Huang, Xu, and Yu, 2015; Ma and Hovy, 2016; Peters et al, 2018]
- Task II: classification (sentence classification OR tagging)
  - Model 1: SVM with features from Stab and Gurevych (2017)
  - Model 2: CNN classifier [Kim, 2014]
  - Model 3 (tagging): CRF-joint (e.g. B-Fact, B-Request, I-Request, etc)
  - Model 4 (tagging): BiLSTM-CRF-joint with ELMo

Experiment

- Segmentation results

	Precision	Recall	F1
FullSent	73.68	56.00	63.64
CRF	66.53	52.92	58.95
BiLSTM + CRF	82.25	79.96	81.09

Experiment

Neural model enhanced with ELMo works the best.

- Segmentation results

	Precision	Recall	F1
FullSent	73.68	56.00	63.64
CRF	66.53	52.92	58.95
BiLSTM + CRF	**82.25**	**79.96**	**81.09**

Experiment

Neural model enhanced with ELMo works the best.

- Segmentation results

	Precision	Recall	F1
FullSent	73.68	56.00	63.64
CRF	66.53	52.92	58.95
BiLSTM + CRF	**82.25**	**79.96**	**81.09**

86.7 on Essays [Stab and Gurevych, 2017]

Experiment

- Classification results

	Overall	Evaluation	Request	Fact	Reference	Quote
Majority	33.30	47.60				
PropLexicon	23.21	22.45	36.07	32.23	59.57	31.28
SVM	51.46	54.05	48.16	52.77	52.27	4.71
CNN	55.48	57.75	53.71	55.19	48.78	33.33
CRF - joint	50.69	46.78	55.74	52.27	55.77	26.47
BiLSTM-CRF - joint	62.64	62.36	67.31	61.86	54.74	37.36

Experiment

- Classification results

Jointly predicting segmentation and type works the best.

	Overall	Evaluation	Request	Fact	Reference	Quote
Majority	33.30	47.60				
PropLexicon	23.21	22.45	36.07	32.23	59.57	31.28
SVM	51.46	54.05	48.16	52.77	52.27	4.71
CNN	55.48	57.75	53.71	55.19	48.78	33.33
CRF - joint	50.69	46.78	55.74	52.27	55.77	26.47
BiLSTM-CRF - joint	**62.64**	**62.36**	**67.31**	**61.86**	**54.74**	**37.36**

Roadmap

- Motivation
- Argument Components
- Annotation
- Experiment
- Analysis
- Conclusion

Analysis

- A larger dataset:
  - OpenReview: ICLR2017, ICLR2018, UAI2018
  - ACL 2017 [Kang, Ammar, Dalvi, van Zuylen, Kohlmeier, Hovy, and Schwartz, 2018]
  - NeurIPS 2013 – 2017 [official website]

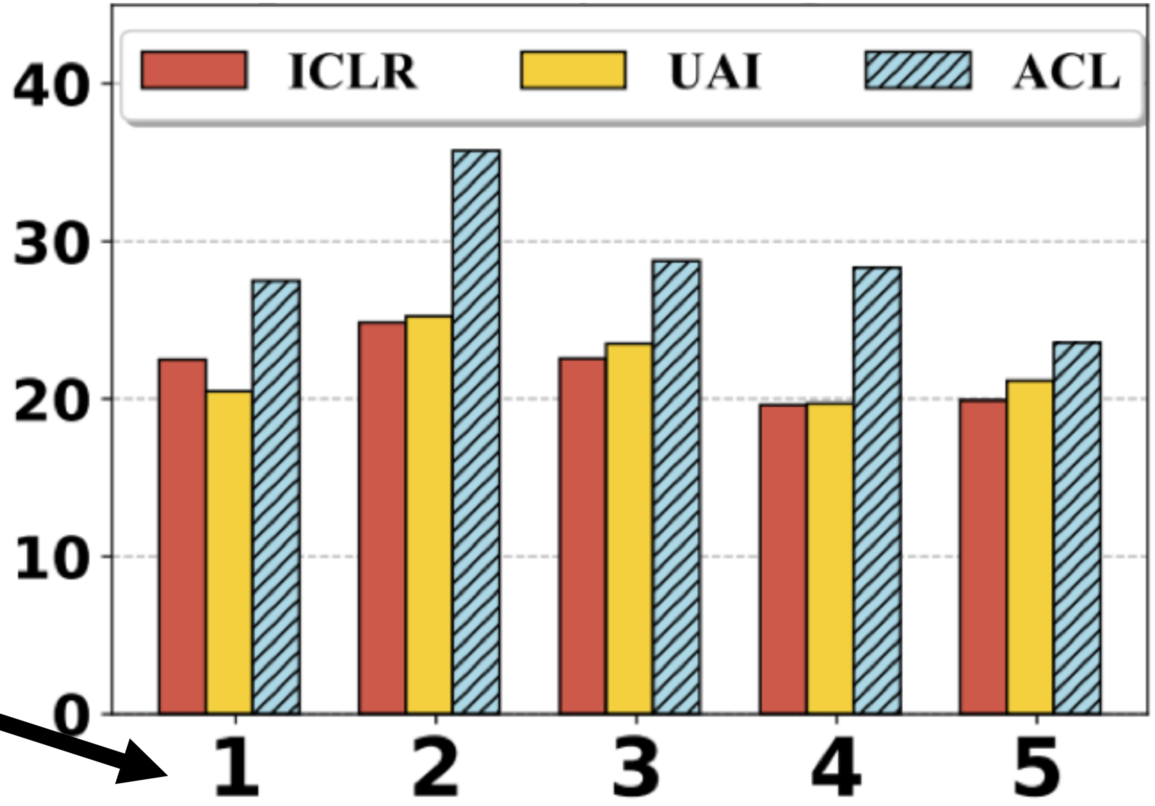
Venue	ICLR	UAI	ACL	NeurIPS	Total
# reviews	4,057	718	275	9,152	14,202

Which venue's reviews contain more arguments?

Which venue's reviews contain more arguments?

- Argument usage by venue and rating

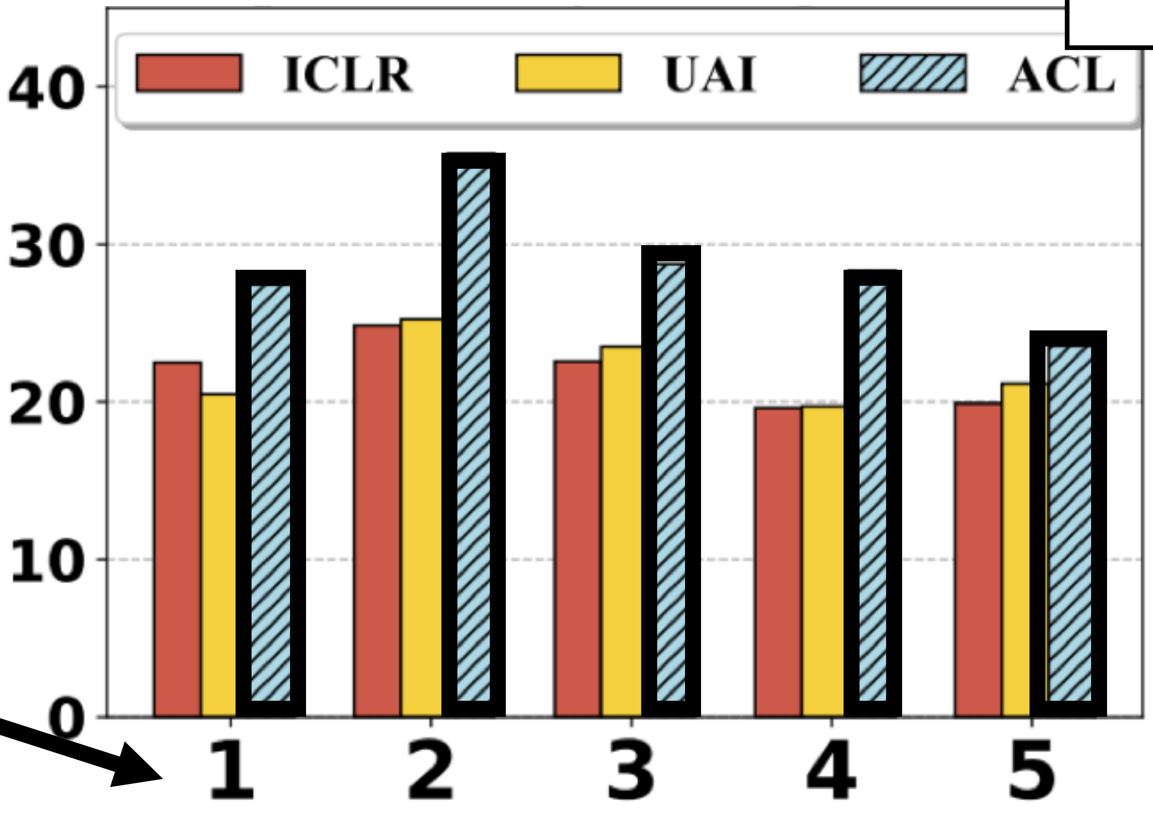
Rating



Which venue's reviews contain more arguments?

- Argument usage by venue and rating

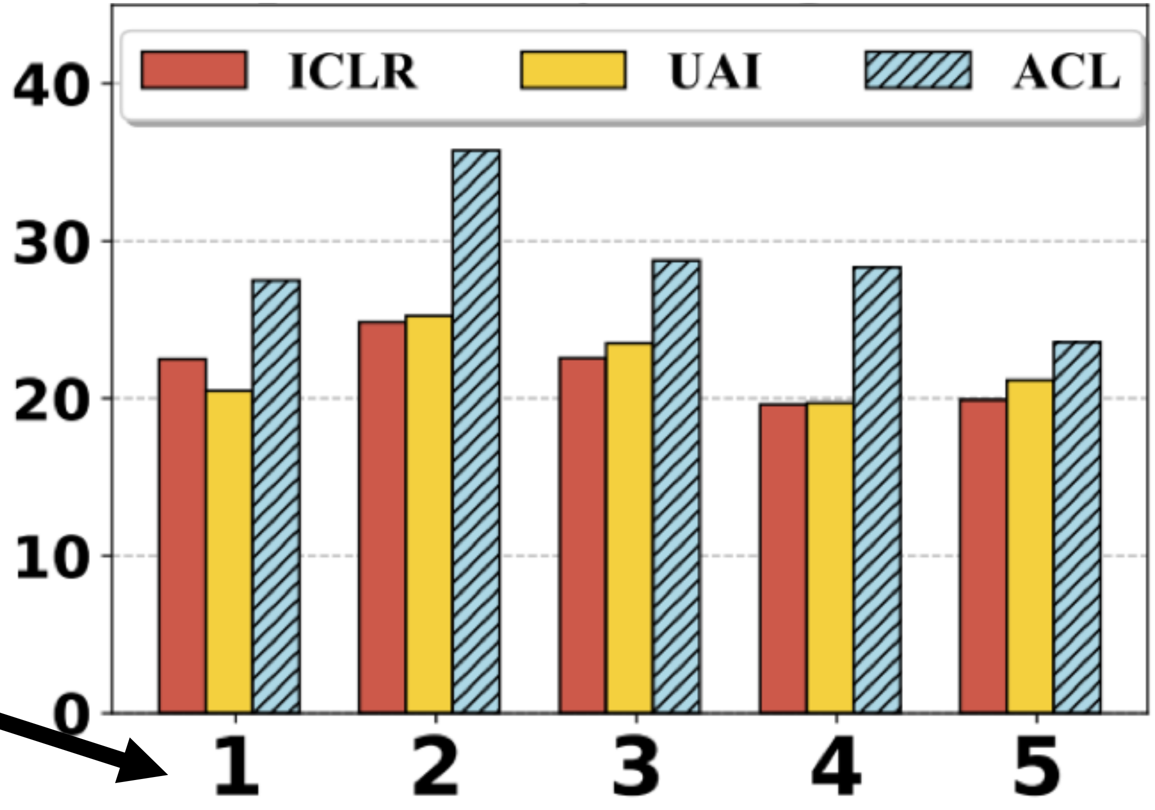
ACL reviews contain more arguments.



Rating

When do the reviewers decide to say more?

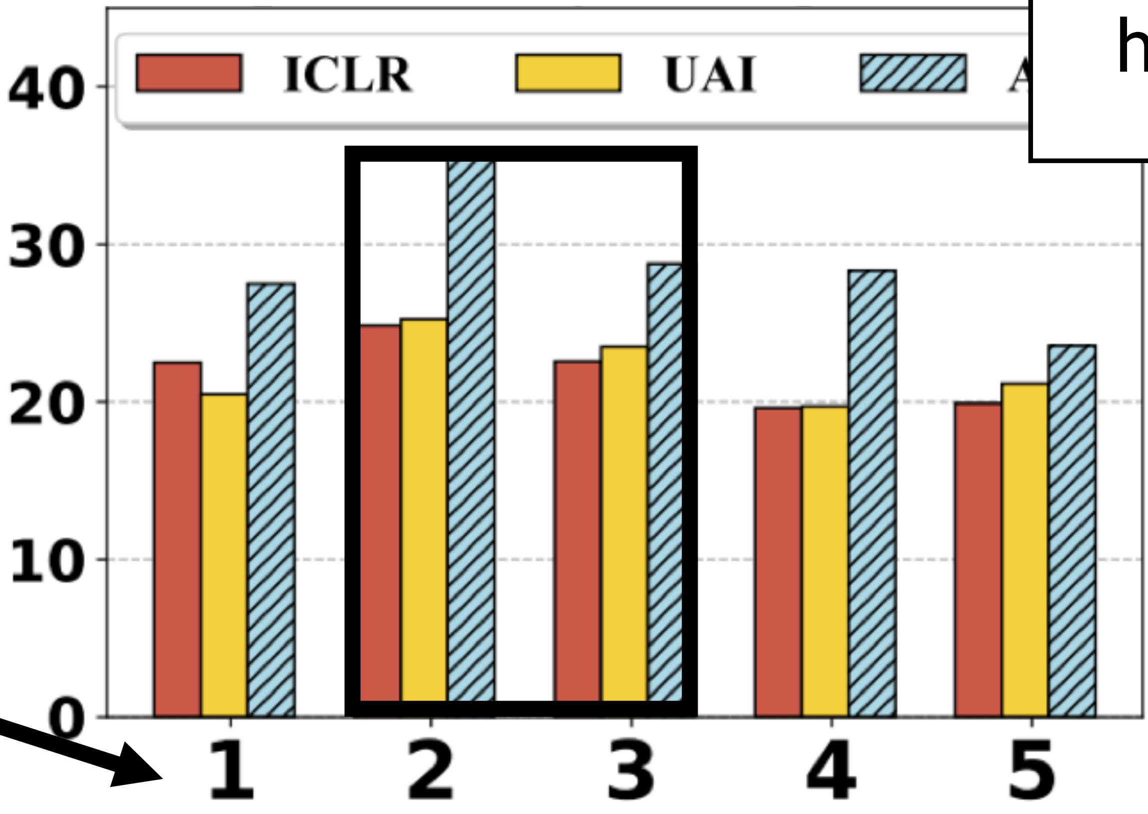
- Argument usage by venue and rating



Rating

When do the reviewers decide to say more?

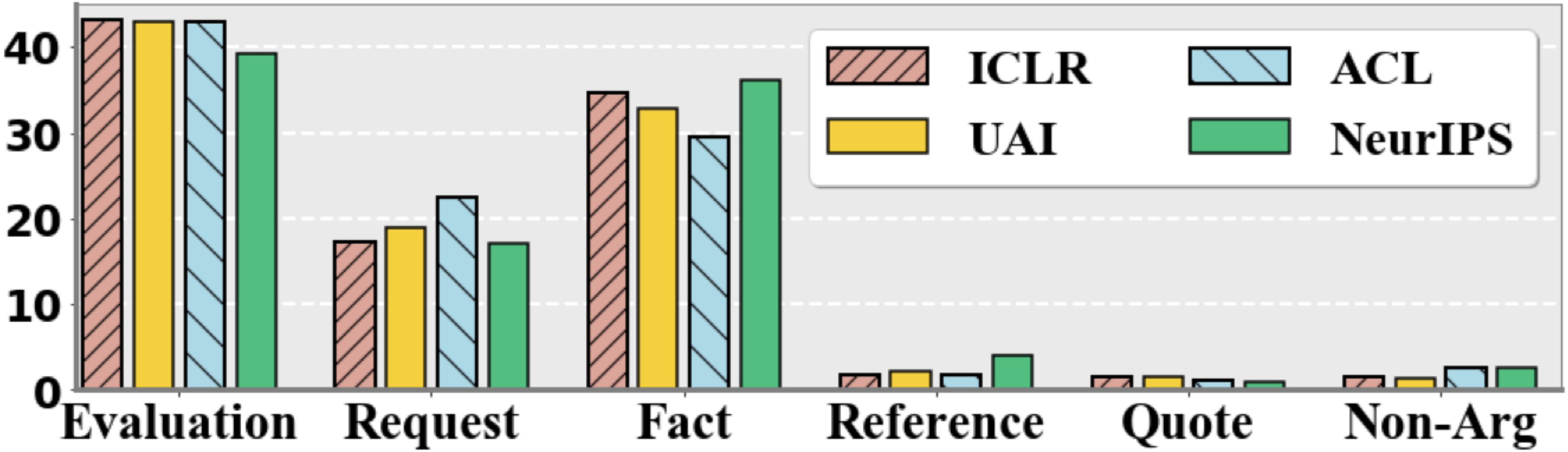
- Argument usage by venue and rating



Borderline reviews have more to say.

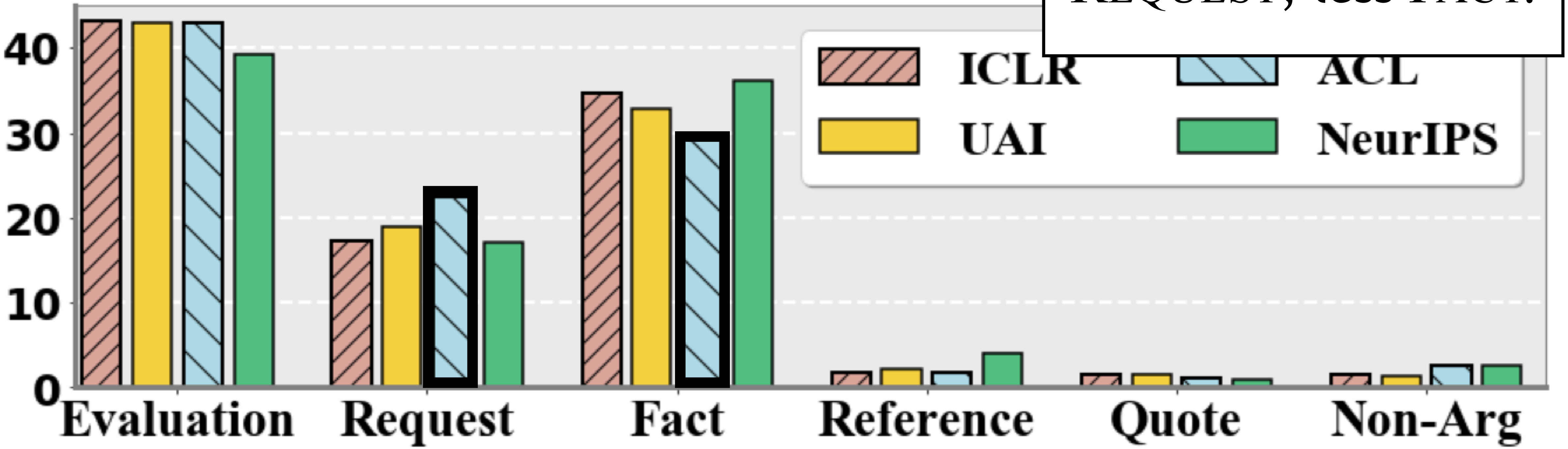
Rating

Does ACL community have preferred argument types?



Does ACL community have preferred argument types?

ACL has more REQUEST, less FACT.



What do reviewers say in each argument type?

- Content in each type (salient words)
  - Method: log-likelihood ratio on term occurrence [Lin and Hovy, 2000]

What do reviewers say in each argument type?

- Content in each type (salient words)

All venues [EVALUATION]

"The experiment section was *unclear*…"

"The contribution of the proposed method is *limited*…"

"The results *seem* unconvincing…"

What do reviewers say in each argument type?

- Content in each type (salient words)

All venues [EVALUATION]

"The experiment section was *unclear*…"

"The contribution of the proposed method is *limited*…"

"The results *seem* unconvincing…"

All venues [REQUEST]

"*Please* also include the majority class baseline"

"The writing *should* be polished"

What do reviewers say in each argument type?

- Content in each type (salient words)

ACL [EVALUATION]

"The paper is well *written*…"

"End-to-end trainable is part of the *strength*…"

"The major *weakness* point is that…"

What do reviewers say in each argument type?

- Content in each type (salient words)

ACL [EVALUATION]

"The paper is well *written*…"

"End-to-end trainable is part of the *strength*…"

"The major *weakness* point is that…"

ACL [REQUEST]

"Please *consider* moving the method to second paragraph"

"Show more *examples*"

What do reviewers say in each argument type?

- Content in each type (salient words)

ACL [EVALUATION]	ICLR [EVALUATION]
"The paper is well *written*…"	"The *network* complexity can …"
"End-to-end …*strength*…"	"The model is *trained* by Adam…"
"The major *weakness* point is that…"	"I am not *convinced* by the experiments…"
ACL [REQUEST]	**ICLR [REQUEST]**
"Please *consider* moving the method …"	"I *recommend* trying a different evaluation…"
"Show more *examples*"	"Showing extra steps in *appendix* …"

Roadmap

- Motivation
- Argument Components
- Annotation
- Experiment
- Analysis
- Conclusion

Conclusion

- We study peer-reviews under an argument mining framework.
- A new review dataset AMPERE (Argument Mining for PEer REview) is annotated for NLP research.
- We employ state-of-the-art methods on a large collection of review dataset, showing distinctive content and argument usage across venues and ratings.

Future Work

- Understand the structures in review arguments
- Design a better data collection method
- Develop tools and interface to improve review quality

Thanks!

- AMPERE dataset, project page: https://xinyuhua.github.io/Resources/naacl19

- An argument mining toolkit will be released soon. Stay tuned!

- Contact: Xinyu Hua (hua.x@husky.neu.edu)

Annotation Difficult Cases

- Difficult cases:
  - Domain specific language:

“The results are significantly better than baselines”

Annotation Difficult Cases

- Difficult cases:
  - Domain specific language:

"The results are significantly better than baselines"

Is it Evaluation or Fact?

Annotation Difficult Cases

- Difficult cases:
  - Domain specific language:

"The results are significantly better than baselines"

Is it Evaluation or Fact?

Annotation Difficult Cases

- Difficult cases:
  - Domain specific language:
  - Rhetorical terms:

"To me, it seems that this is a causal model with a neural network (NN) modeling..."

Annotation Difficult Cases

- Difficult cases:
  - Domain specific language:
  - Rhetorical terms:

"To me, it seems that this is a causal model with a neural network (NN) modeling…"

Is it Evaluation or Fact?

Annotation Difficult Cases

- Difficult cases:
  - Domain specific language:
  - Rhetorical terms:

"To me, it seems that this is a causal model with a neural network (NN) modeling..."

Is it Evaluation or Fact?

Annotation Difficult Cases

- Difficult cases:
  - Domain specific language:
  - Rhetorical terms:
  - Questions:

"What is the number of parameters?"

Is it Evaluation or Non-Arg?

Annotation Difficult Cases

- Difficult cases:
  - Domain specific language:
  - Rhetorical terms:
  - Questions:

"What is the number of parameters?"

Is it Evaluation or Non-Arg?

Annotation Difficult Cases

- Difficult cases:
  - Domain specific language:
  - Rhetorical terms:
  - Questions:

"How could the number of parameters be this large?"

Annotation Difficult Cases

- Difficult cases:
  - Domain specific language:
  - Rhetorical terms:
  - Questions:

"How could the number of parameters be this large?"

Is it Evaluation or Non-Arg?

Annotation Difficult Cases

- Difficult cases:
  - Domain specific language:
  - Rhetorical terms:
  - Questions:

"How could the number of parameters be this large?"

Is it Evaluation or Non-Arg?