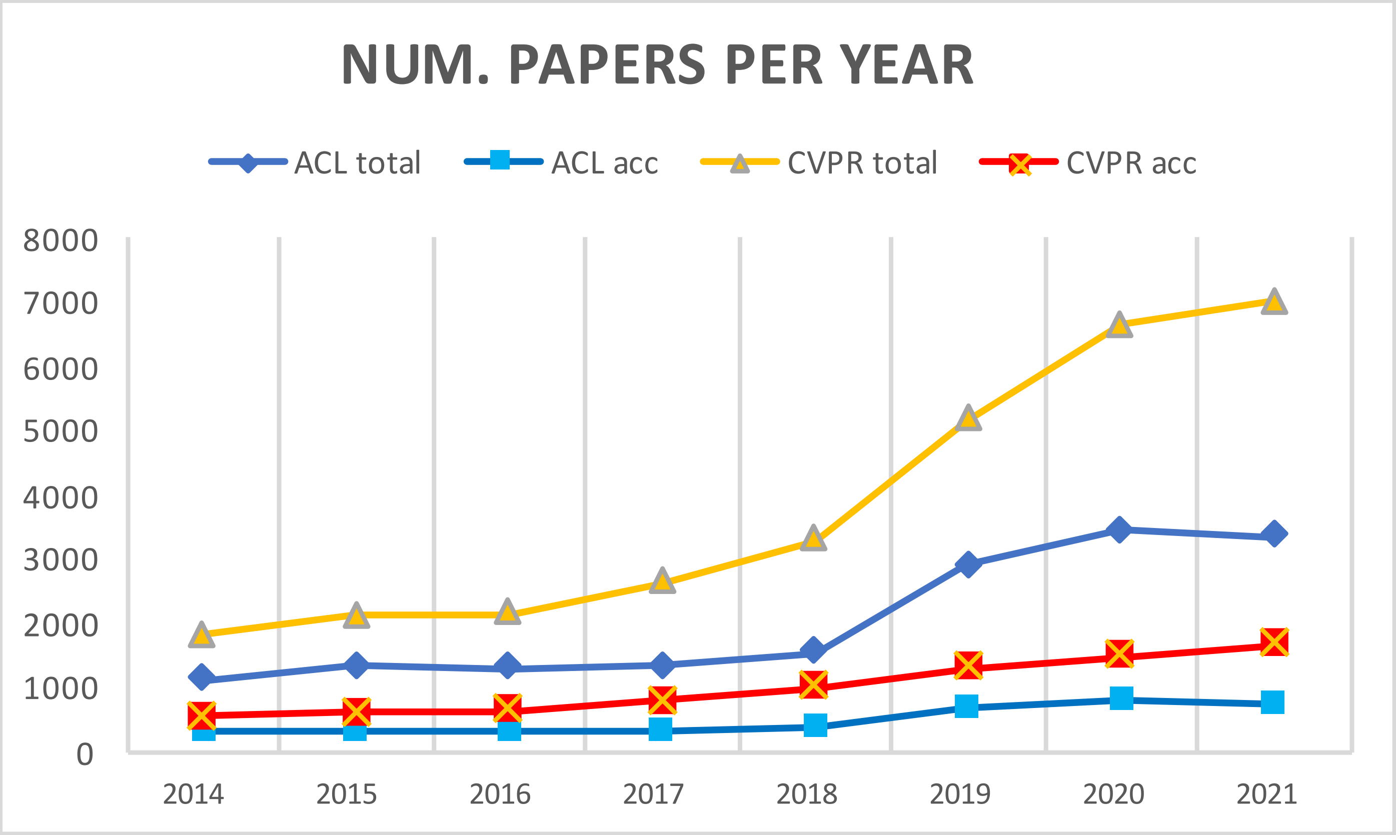
What do writing features tell us about AI papers?

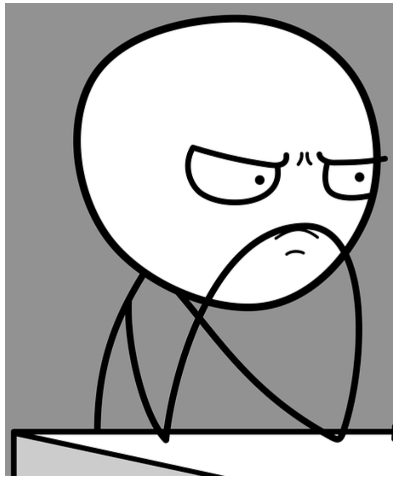
Zining Zhu, Bai Li, Yang Xu, Frank Rudzicz

Recent submissions increase in numbers



Data source: https://github.com/lixin4ever/Conference-Acceptance-Rate

Problems from two sides



Poorly organized

Methodology is problematic

Result is unclear

Question - analysis mismatch

Limited novelty

Limited impact

Ethical concerns

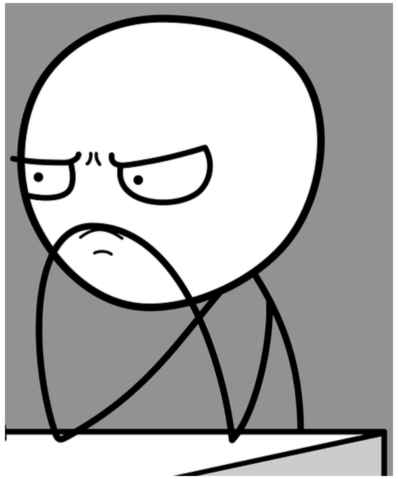
Didn’t read carefully

Doesn’t understand our method

Doesn’t think hard

Doesn’t understand the field

Reviewers are paranoid

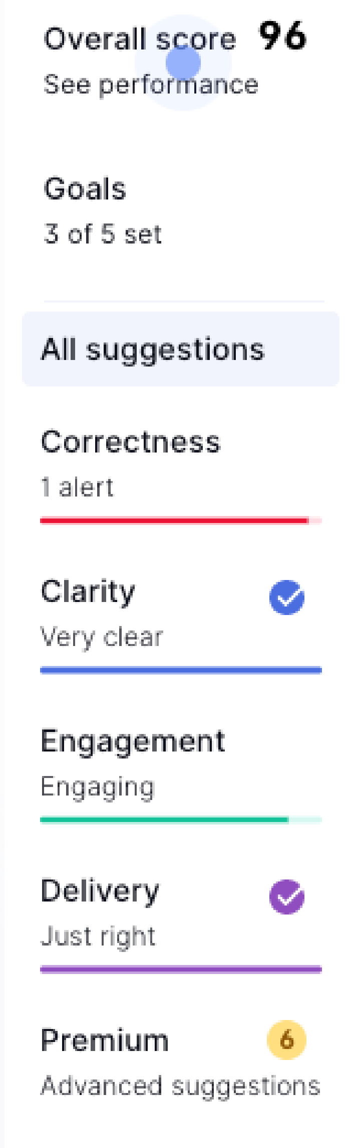


Gets random submissions

Gets random peer reviews

Possibilities of improvements?

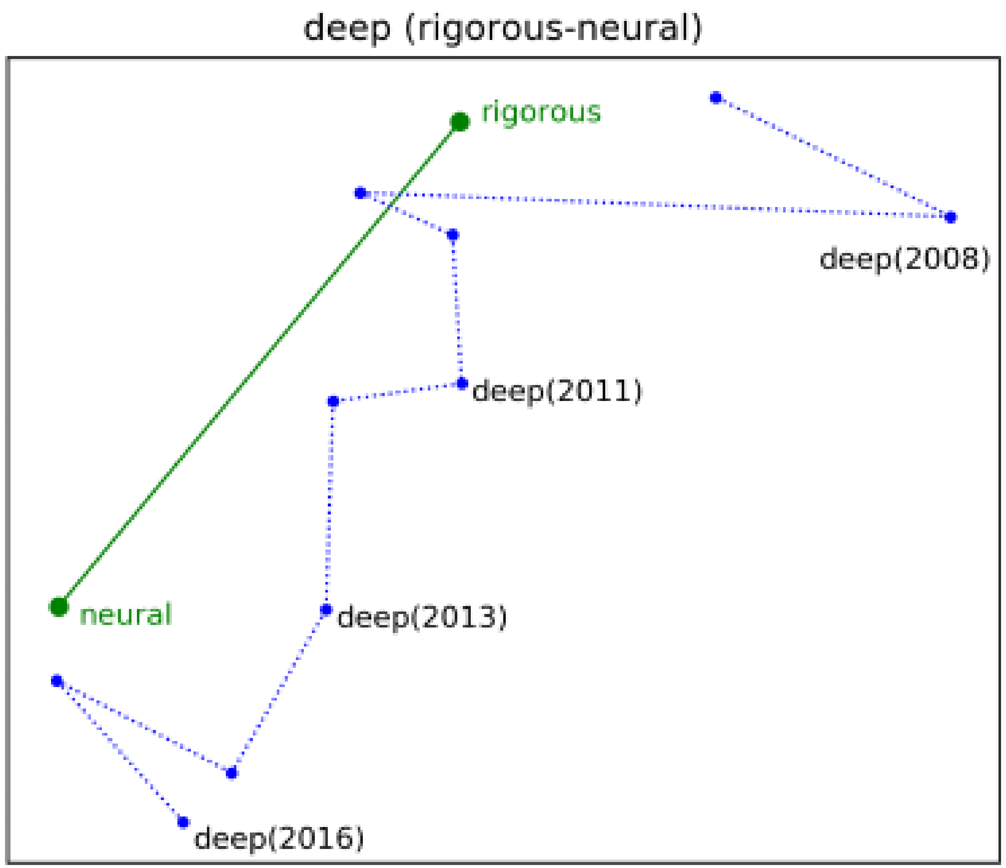
- Improved peer review procedure
  - OpenReview
  - ACL Rolling Review
- Use DNN to predict paper outcomes
  - Text classification problem
- Intuition: text markers can lead to scalable solutions
  - “Best of both worlds”
  - Similar: Automatic Essay Scoring, e.g., Grammarly →



There are some interesting text markers

An example: locations on the *semantic coordinates*.

- Hypothesis: word semantics shift along certain coordinates.
- Semantically stable words form coordinates.
- Target words shift along the coordinates.

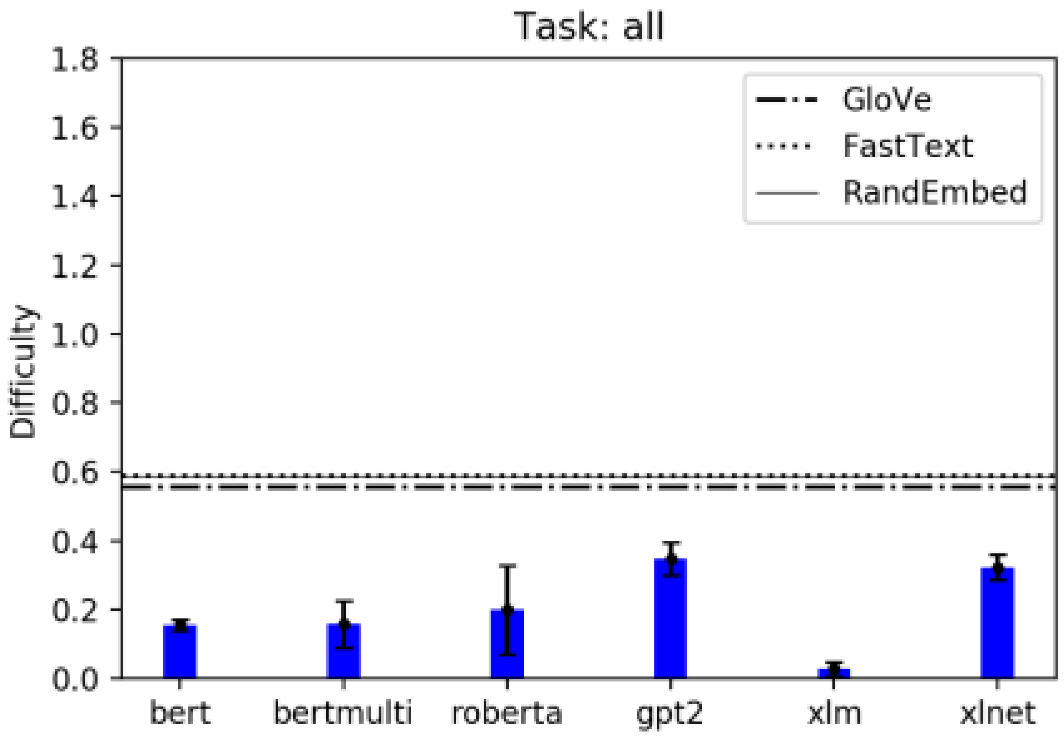


Zhu, Z., Xu, Y., & Rudzicz, F. (2020). Semantic coordinates analysis reveal language changes in AI research. *ArXiv:2011.00543*. https://arxiv.org/abs/2011.00543

And they can be useful

An example: examine the rhetorical capacities of neural LMs.

- Use simple models ("probes") to predict rhetorical features
- Use loss to measure the difficulty → negation of "the goodness of encoding the knowledge"



Zhu, Z., Pan, C., Abdalla, M., & Rudzicz, F. (2020). Examining the rhetorical capacities of neural language models. [EMNLP BlackboxNLP]

Let’s try some text markers for AI papers

We consider 74 *writing features*

i.e., do **not** explicitly describe the semantics.

- Metadata: outbound citations, article lengths, sentence lengths…
- Readability: Flesch, Flesch-Kincaid, semantic surprisal
- Lexical richness: Moving-average type-token ratio
- Syntactic: Grammar error counts, active / passive voice portions
- Stylistic features: POS signal constituency, RST signal constituency

Zhu, Z., Li, B., Xu, Y., & Rudzicz, F. What do writing features tell us about AI papers? *Submitted (2021)*

What can the writing features do?

- They are correlated to Conference (C) vs Workshop (W) appearance.

Venue	Features	Spearman R	ATE	Interpretation
ACL	flesch_kincaid_grade_level_bodytext	−0.05	+0.05	Ambiguous
	grammar_errors_abstract	−0.09**	−0.01	W papers are larger
	surprisal_abstract_std	−0.01	+0.00	Ambiguous
	title_word_length	−0.09**	−0.01	W papers are larger
	voice_bodytext_active	+0.09**	+0.15	C papers are larger
EMNLP	outbound_citations_per_word	−0.17**	+67.6	Ambiguous
	n_author	−0.17**	−0.05	W papers are larger
	grammar_errors_abstract	−0.18**	+0.01	W papers are larger
	n_outbound_citations	−0.09	+0.09	Ambiguous
	abstract_word_counts	−0.16**	+0.00	W papers are larger

Zhu, Z., Li, B., Xu, Y., & Rudzicz, F. What do writing features tell us about AI papers? *Submitted (2021)*

What can the writing features do?

- They can predict Conference (C) vs Workshop (W) appearance.

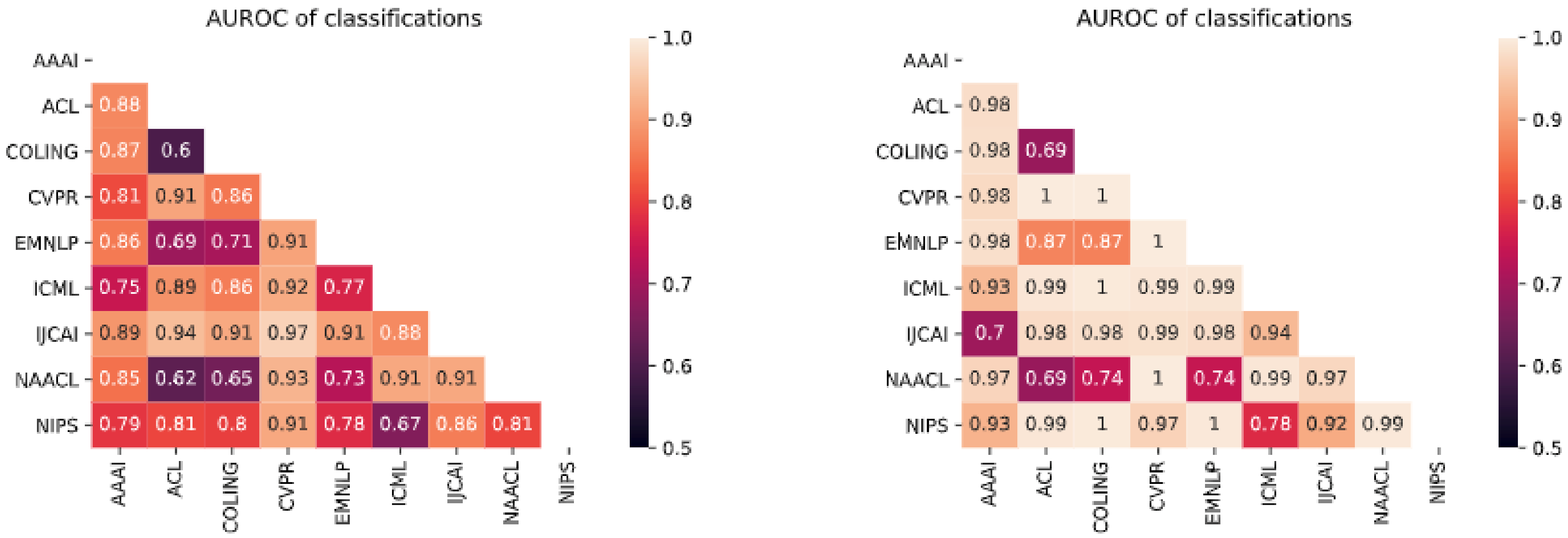
Venue Name		Writing Features					TF-IDF		RoBERTa
	74 features	RST	Surprisal	Grammar	LexRich	Readability	Full text	Abstract	Abstract
AAAI	.755(.028)	+.001	+.024	+.010	−.002	+.009	$+.206^{**}$	$+.203^{**}$	$+.212^{**}$
ACL	.867(.004)	+.001	+.000	+.001	+.001	+.001	−.004	$-.008^{*}$	−.015
COLING	.837(.010)	+.005	+.005	+.003	+.003	+.004	$+.049^{**}$	$+.051^{**}$	$+.052^{**}$
CVPR	.900(.005)	−.007	−.006	−.001	−.006	−.005	$-.052^{**}$	$-.067^{**}$	$-.070^{**}$
EMNLP	.737(.020)	+.003	+.014	+.012	+.022	+.015	$+.159^{**}$	$+.153^{**}$	+.102
ICML	.659(.023)	$-.277^{**}$	$-.042^{*}$	−.102	$-.262^{**}$	−.185	$+.333^{**}$	$+.333^{**}$	$+.300^{**}$
IJCAI	.868(.002)	$-.067^{**}$	$-.066^{**}$	$-.045^{**}$	$+.075^{**}$	$-.067^{**}$	$-.029^{**}$	$-.061^{**}$	−.076
NAACL	.757(.019)	+.016	+.016	+.011	$+.017^{*}$	$+.016^{*}$	$-.107^{**}$	$-.128^{**}$	−.182
NeurIPS	.586(.035)	$-.193^{**}$	−.039	−.097	$-.212^{**}$	−.157	+.031	$-.077^{*}$	−.110

Table 3: F1 scores of the C vs. W classification results. The second column show
using 74 writing features. The remaining columns show the values *relative to* th
$p < .005$ and $p < .001$ respectively, both on 2-tailed t-test with $dof = 10$, Bon

Sometimes comparable to TF-IDF features, and even RoBERTa.

What can the writing features do?

- They can sort of tell apart between different venues.



(a) With all writing features. (b) With tf-idf features.

Figure 1: The AUROC of inter-venue classifications. The venues in the same categories (e.g., COLING and ACL) are harder to tell apart than other venues, using either the writing features or tf-idf features.

What can the writing features do?

- They can not predict the inbound citation counts.

Venue Name	Writing Features						Baseline	TF-IDF	
	74 features	RST	Surprisal	Grammar	LexRich	Readability		Full text	Abstract
AAAI	+8.61	+0.09	+0.69	+0.27	+1.16	+0.61	20.77(27)	0.07(.02)	0.07(.01)
ACL	+6.27	+0.48	+0.11	+0.10	+0.34	+2.89	389.76(636)	0.15(.01)	0.17(.01)
COLING	+248.87	+3.51	+0.05	+0.32	+0.62	+368.18	437.76(1006)	0.15(.02)	0.16(.01)
CVPR	−6.11	+239.82	+9.62	+6.90	+22.24	+12.35	15273.45(24710)	0.17(.01)	0.19(.01)
EMNLP	+55211.48	+8.12	+4.66	+42.06	+12.11	+458.65	1194.59(2788)	0.15(.02)	0.17(.03)
ICML	+45.13	+6.93	+37.59	+9.77	+988.27	+86.97	1279.15(1200)	0.02(.01)	0.02(.02)
IJCAI	+8.84	+1.37	+0.81	+1.20	+3.77	+1.97	23.77(25)	0.16(.01)	0.22(.04)
NAACL	+18.04	+2.22	+0.83	+0.99	+0.08	+195.92	420.34(855)	0.22(.01)	0.22(.01)
NeurIPS	+78.25	+4.64	+6.37	+39.81	−2.87	+76.99	3305.99(5216)	0.20(.01)	0.23(.01)

But TF-IDF features can predict!

More about the data

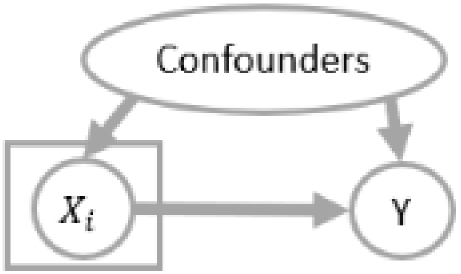
- Computed features on 945,674 CompSci articles from S2ORC.
  - 97.68% have ≤10 annual income citations.
  - Each article is cited 1.59 (std=13.5) times per year.
- Gave C & W labels for AI venues.
  - NLP: ACL, COLING, EMNLP, NAACL
  - AI: AAAI, IJCAI
  - ML: ICML, NeurIPS
  - CV: CVPR
  - ICRA and ICASSP not used

Venue Name	N. articles	N. articles by label C	N. articles by label W
AAAI	624	395	229
ACL	2,836	2,175	661
COLING	1,860	1,353	507
CVPR	3,495	2,824	671
EMNLP	714	437	277
ICML	930	396	534
ICRA	703	662	41
IJCAI	632	423	209
NAACL	2,142	1,354	788
NeurIPS	930	396	534

Table 7: Number of `C` and `W` articles of each venue. The arXiv papers of the corresponding sections are included as `W` papers. For example, `cs.Learning` and `cs.ML` are included in the `W` portions of ICML and NeurIPS.

More about the writing features…

- They are mutually dependent
  - Causal model assumed independence -> Observe multicollinearity effect.
  - Partial features can often predict well.



- They describe more than "just the writing".
  - E.g., RST: stylistic choices -> author -> content
  - E.g., title length -> scope of content -> num. readers -> citation counts
- BTW: Good papers are more than well-"written".
  - Should consider their impact.

Summary

- Computed 74 *writing features*
- Compiled a test suite to assess their usefulness:
  - Conference vs. Workshop appearance prediction
  - Venue appearance prediction
  - Citation counts prediction
- Text markers can lead to scalable, high-quality, and trustworthy solutions for assessing academic article writing.
  - More text markers, and group them together.
  - Additional subjects, more than just CompSci / AI

Zhu, Z., Li, B., Xu, Y., & Rudzicz, F. What do writing features tell us about AI papers? *Submitted (2021)*

Connections beyond academic writing

Structural probe, edge probing, and rediscovering classic NLP pipeline

Diagnostic classification

How is BERT surprised? (Li et al., 2021)

What does BERT learn about the structure of language? (Jawahar et al 2019)

Linguistic theory

Interpretable text-markers

Symbolic reasoning

Predicting the inductive bias (Lovering et al., 2021)

RoBERTa acquires linguistic features eventually (Warstadt et al., 2020)

ML4H

Lexical features are more vulnerable (Balagopalan et al., 2019)

Agreeing on interpretations of linguistic features (Zhu et al., 2019)