A Dataset of Peer Reviews (PeerRead)

Collection, Insights and NLP Applications

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi Mishra, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, Roy Schwartz

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Carnegie Mellon University
Language Technologies Institute

UW NLP

1. Summary

Example peer review:

This paper details the approach that won the … competition … an approach that predicts … The approach is a collection of different methods, but it yields impressive empirical results, and it is a clear, well-written paper.

Annotations:

Aspect	Score (1-5)
Impact	4
Originality	3
.....	
Clarity	5

Motivation:

- ❑ Enable scientific study of the peer-review process: **consistency**, **bias**, **review quality, etc.**
- ❑ Automated tools to assist authors, reviewers and area chairs

Contributions:

- ❑ The **first public dataset of scientific peer reviews**:
  - ❑ **14.7K papers** with **accept/reject** decisions and **10.7K textual peer-reviews**
- ❑ Data analysis reveals interesting phenomena in the peer reviews
- ❑ Two new NLP tasks to promote research in this area
  - ❑ Baseline models substantially outperform majority baselines

2. Dataset

- ❑ Accept/Reject annotations
- ❑ Aspect score annotations
  - ❑ 1.3K ICLR 2017 reviews manually annotated with aspect scores

Section	#Papers	#Reviews	Asp.	Acc / Rej
NIPS 2013–2017	2,420	9,152	×	2,420 / 0
ICLR 2017	427	1,304	✓	172 / 255
ACL 2017	137	275	✓	88 / 49
CoNLL 2016	22	39	✓	11 / 11
arXiv 2007–2017	11,778	—	—	2,891 / 8,887
total	14,784	10,770		

Analysis

Aspect	ρ
Substance	0.59
Clarity	0.42
Appropriateness	0.30
Impact	0.16
Meaningful comparison	0.15
Originality	0.08
Soundness/Correctness	0.01

Aspects in descending order of their **correlation with acceptance**

Presentation format	Oral	Poster
Recommendation	3.83	2.92
Substance	3.91	3.29
Clarity	4.19	3.72
Meaningful comparison	3.60	3.36
Impact	3.27	3.09
Originality	3.91	3.88
Soundness/Correctness	3.93	4.18

Mean Aspect score values for papers accepted with **oral/poster** presentations

3. NLP TASK: Paper Acceptance Classification

Task: Given a paper text, predict whether it will get accepted to one of our target conferences

- ❑ NLP (*ACL, EMNLP), ML (ICML and NIPS) and AI (AAAI)
- ❑ Features
  - ❑ **Coarse features** (e.g. title length, whether terms such as '*neural*' appear in the abstract...)
  - ❑ **Lexical features** (e.g., CBOW, N-grams, GloVe embeddings…)
- ❑ Model: We explored several off-the-shelf classifiers (e.g., SVM, KNN)

	ICLR	cs.cl	cs.lg	cs.ai
Majority	57.6	68.9	67.9	92.1
Ours	65.3	75.7	70.7	92.6
(Δ)	+7.7	+6.8	+2.8	+0.5

Accept/reject classification accuracy

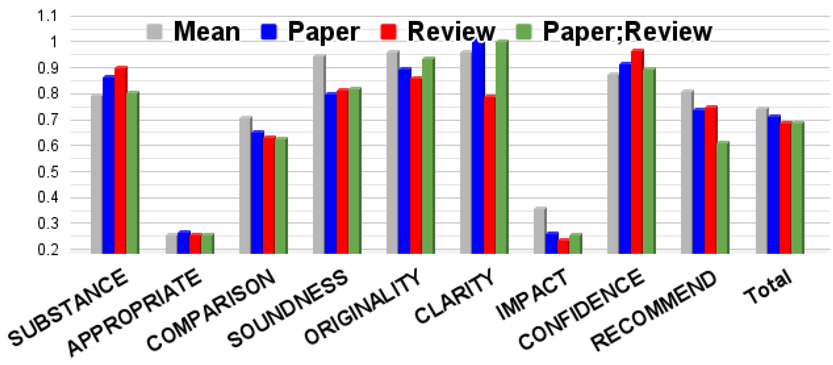
Best model	65.3
– appendix	–5.4
– num_theorems	–3.8
– num_equations	–3.8
– avg_len_ref	–3.8
– $\text{abstract}_{\text{state-of-the-art}}$	–3.5
– #recent_refs	–2.5

Feature ablation

4. NLP TASK: Review Aspect Score Prediction

Task: Predict the numerical values for aspect scores given the paper and review text

- ❑ Our model: text encoder (CNN, LSTM, DAN). Baseline: **Mean** aspect score



Aspect score prediction using different inputs (mean-squared error loss, **lower is better**)

Dataset and code available at https://github.com/allenai/PeerRead