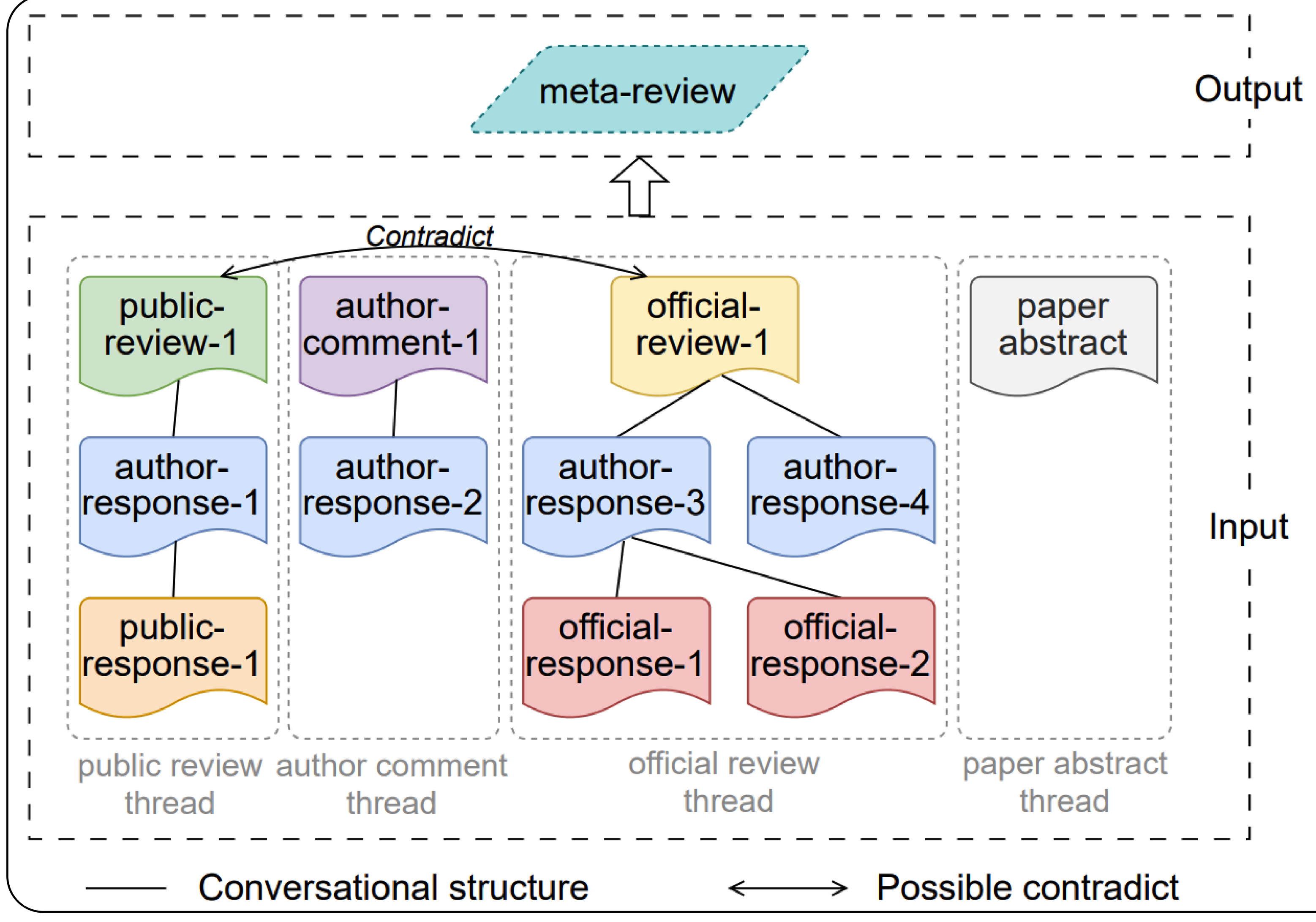
THE UNIVERSITY OF MELBOURNE

Summarizing Multiple Documents with Conversational Structure for Meta-Review Generation

Miao Li[#], Eduard Hovy[#*], Jey Han Lau[#]
[#]The University of Melbourne [*]Carnegie Mellon University
miao4@student.unimelb.edu.au, eduard.hovy@unimelb.edu.au, laujh@unimelb.edu.au

Meta-review Generation -> Multi-Document Summarization



- Meta-reviewers need to comprehend and carefully summarize information from individual reviews, multi-turn discussions between authors and reviewers and the paper abstract in practice
- We formulate the creation of meta-reviews as an abstractive multi-document summarization (MDS) task
- Most content of meta-reviews can be anchored to source documents in samples both with and without conflicts
  - Word-level human annotation
  - In CF samples, at least two reviewers have very different scores (≥ 4)

Data	#Samples	Mean Rating Variance	Anchored Words
Non-CF (w/o conflicts)	25	0.717	79.67%
CF (w/ conflicts)	35	6.668	72.74%

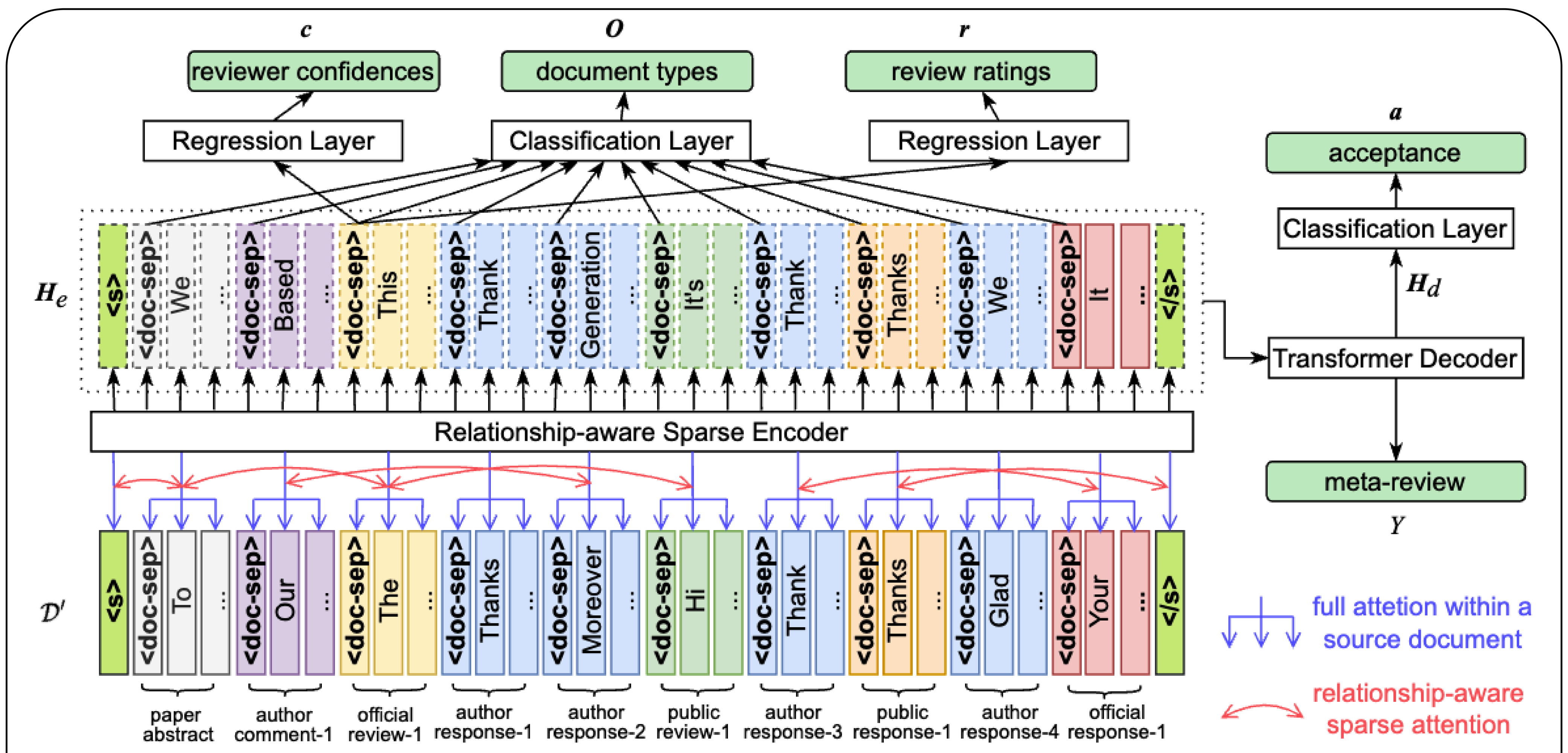
The Constructed PeerSum Dataset (11,995/1,499/1,499)

- Meta-reviews are largely **faithful** to the corresponding source documents despite being highly abstractive in novel n-grams
- Source documents have rich inter-document relationships with **an explicit conversational structure**
- Source documents occasionally feature ***conflicts*** (13.6% samples with conflicts)
- There is a rich set of **metadata**, such as document type, review rating/confidence and paper acceptance outcome
- Paper acceptance is used to assess the quality of automatically generated meta-reviews (**the newly proposed evaluation metric**)

"The main contribution is the novel pre-training strategy introduced. The work has potential high impact in the research area..." VS *"The approach proposed in the paper seems to be a small incremental change on top of the previous GNN pre-train work. The novelty aspect is low."*

Introduction section is not well-written. VS *This paper is well written and looks correct.*

The RAMMER Model



Encoder Auxiliary Loss

$$\mathcal{L} = \alpha_g \mathcal{L}_g + \alpha_c \mathcal{L}_c + \alpha_r \mathcal{L}_r + \alpha_o \mathcal{L}_o + \alpha_a \mathcal{L}_a \qquad \boldsymbol{H}_l = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T \odot \sum_j \beta_j \cdot \boldsymbol{R}_j^{\dagger}}{\sqrt{d_k}}\right)\boldsymbol{V}$$

Cross Entropy   Decoder Auxiliary Loss

- Developed sparse attention for pre-trained encoder-decoder models to capture the conversational structure of source documents
  - Different attention heads pay different attention on relationships derived from the tree-like conversational structure
  - Expected to learn sematic relationships with the help of recognition of the conversational structure
- Trained with auxiliary objectives which are to predict metadata such as review rating and confidence, and the paper acceptance

Experiments (Automatic Evaluation)

Model(#Params)	Test Data	R-L↑	BERTS↑	ACC↑
PEGASUS (568M)	Non−CF	27.24	14.75	0.725
PRIMERA (447M)	Non−CF	28.70	12.67	0.725
LED (459M)	Non−CF	29.52	16.59	0.748
PegasusX (568M)	Non−CF	29.65	17.36	0.745
RAMMER (459M)	Non−CF	**30.39***	**17.42***	**0.768**
PEGASUS (568M)	CF	26.77	13.66	0.649
PRIMERA (447M)	CF	29.13	12.33	0.639
LED (459M)	CF	29.19	15.32	0.698
PegasusX (568M)	CF	**29.30**	15.69	0.707
RAMMER (459M)	CF	29.19	**15.88***	**0.724**

- Our RAMMER performs better than other models, especially in predicting the paper acceptance (ACC)

Experiments (Human Evaluation)

- Models mostly fail to recognize (i.e., identifying conflicting information) and resolve (i.e., reaching similar final decision to the human meta-reviewer) conflicts in its meta-reviews (40 samples)

Model	Recognition	Resolution
PRIMERA	3/23	2/23
LED	4/23	4/23
PegasusX	5/23	5/23
RAMMER	8/23	3/23

The 2023 Conference on Empirical Methods in Natural Language Processing, December 6 –10, Singapore