Response to Reviewer 1

Q1: Notations and abbreviations.
A1: We will use O^e and I^e to denote the set of OAs and ISs extracted from all reviews of an entity e.
Q2: Illustrations of a running example and different models.
A2: We will change Figure 1 in paper to a running example and add the illustration of BAG and BAI to Figure 2 in paper.
Q3: Evaluation: Human evaluation.
A3: In Section 3.4, we briefly described the human evaluation as "considering the consistency with multi-review and informativeness together". The detailed instruction for annotators were: Consistency: how well the sentiment of summary agrees with multi-review? Informativeness: how much useful information does the summary provide?

To enhance human evaluation, we ask three annotators to evaluate summaries under 5 aspects: Fluency (Flu.), Coherence (Coh.), Non-redundancy (NR.), Consistency (Cons.) and Overall. The new results on Best-worst scaling (Section 3.4) are shown in Table 1.

Data	Model	Flu.	Coh.	NR.	Cons.	Overall	AC
Yelp	Gold	0.34	0.49	0.41	0.35	0.31	-
	TransSum	-0.46	-0.53	-0.70	-0.64	-0.48	0.36
	MB-B	0.12	0.14	0.29	0.29	0.17	0.42
Amazon	Gold	0.32	0.55	0.38	0.44	0.32	-
	TransSum	-0.54	-0.67	-0.68	-0.72	-0.41	0.32
	MB-B	0.22	0.12	0.30	0.28	0.09	0.38

Table 1: Human evaluation.

Q4: Evaluation: Aspect coverage (AC).
A4: To show the effectiveness of rule-based aspect extraction method in evaluation, we ask three annotators to extract aspects from reference summaries and computed ROUGE-1 (R-1) recall between manual extracted and rule-based aspects. The recall of Amazon is 84% and Yelp is 88%, which shows the rule-based method can extract most aspects.

We agree that the automatically extracted aspects may be not good enough for evaluation. Given a generated summary and its reference summary, we ask annotators to extract their aspects and compute R-1 recall between the aspects from generated summary and reference summary. The AC in Table 1 shows the average of R-1 recall from three annotators.

Model		Yelp			Amazon		
		R-1	R-2	R-L	R-1	R-2	R-L
BART	EM OAs	33.22	5.54	16.83	30.16	5.24	17.95
	MM OAs	20.32	3.24	11.14	19.67	2.97	9.32
	BAG	34.79	6.83	18.42	31.27	6.31	19.14

Table 2: Ablation study about EM OAs and MM OAs.

Q5: Evaluation: Ablation study.
A5: Table 7 in paper is our ablation study, which compares the different ways of using noisy OAs and ISs. According to the comments, we add the ablation study about the components of noisy OAs, i.e., EM OAs and MM OAs. Here we just show the BART-based (best) results. BAG performs best in Table 2. We can't use only OAs or ISs of summary as input since the summary is not given during test.
Q6: Pretrained LM.
A6: PlanSum, TransSum and our MB-B use different pretrained LMs. MB-B can finetune the weights of pretrained LM, which is difficult for other two methods because of their model design. It is unfair to classify them into the methods without pretrained LM, so we no longer classify the methods. We will analyze their usage of pretrained LMs.
Q7: Rule-based MIN-MINER and syntactic rules.
A7: As OAs extraction is not our contribution, we provided the citation about MIN-MINER and syntactic rules. The use of OA extraction is flexible. The better the extraction method used, the better the results of our approach.

Response to Reviewer 2

Q1: ROUGE.
A1: The ROUGE scores in Table 3 are the published version. In our baselines, Copycat and FewSum used google rouge. As the test set of FewSum is different from other baselines (See footnote 5 and 6 in paper), we retested FewSum and evaluated it by pyrouge. The ROUGE scores (R-1/R-2/R-L) of Copycat based on pyrouge are 28.95/4.80/17.76 (Yelp) and 31.84/5.79/20.00 (Amazon), which are smaller than their published version. We will provide the ROUGE scores of all methods on these two tools in the appendix.
Q2: Human evaluation.
A2: According to the suggestion, we follow Copycat for human evaluation (Refer to A3 in Response to Reviewer 1).
Q3: Deeper semantic-level similarities make summaries more abstractive.
A3: We agree with this. But such methods will bring more noise. For example, OA pairs, which are consistent in sentiment but different in aspects, will also have high similarity.
Q4: Parameters.
A4: The initialized parameters that are in MAI but not in BAG is the parameters of BART-large. The parameters of two encoders are not shared.
Q4: Aggregate the information of OAs using LSTM.
A4: We added special tokens before each OA pair and took special tokens in different positions to denote different OAs. So we used LSTM to aggregate all special tokens to represent the set of OAs rather than one special initial token.

Response to Reviewer 3

Q1: Related work.
A1: In Related work, these work (Elsahar et al. 2021; Tian, Yu, and Jiang 2019; Gerani et al. 2014; Ma et al. 2020; Fabbri et al. 2019) are not mentioned in Section 1. We will compress the description of the work introduced in Section 1.
Q2: Example in Table 8.
A2: We will add more analysis as follows: "Compared with gold summary, the sentence in generated summaries are not coherent enough and the coreference is not clear. For example, the sentences in MB output on food were incoherent and the 'it' in the last sentence denotes the restaurant. The last sentence of MB output is inferred by adding ISs."
Q3: Section 2 and other typos.
A3: We will revise the Section 2 (Refer to A1 and A2 in Response to Reviewer 1) and correct the typos.

Response to Reviewer 4

Q1: Novalty.
A1: As far as we know, we are the first to build noisy OAs and noisy ISs according to OAs and ISs in the sampled summary to simulate multi-reviews.
Q2: Conflicting views.
A2: In our approach, we first sample a review as summary and extract its OAs and ISs. Then we extract the OAs and ISs from the other reviews. We construct the noisy OAs and ISs as input according to the similarity between the OAs and ISs from other reivews and sampled summary, which ensures that the the noisy OAs and ISs is consistent with the summary. Besides, it is rare that the sampled summary is inconsistent with all other reviews.
Q3: Implicit information.
A3: Most sentences in the reviews are always short. We compute the percentage of length of the sentences removed OAs and stopwords from its original length. The Yelp is about 10.7% and Amazon is about 11.1%. We also sample 100 reviews from Yelp and Amazon respectively and ask human annotators to record the sentences with loss of information after removed OAs. The percentage of such sentences are 9.1% (Yelp) and 9.3% (Amazon). These results show that the sentences from which OAs can be extracted contain very little implicit information.
Q4: Test samples.
A4: According to the comment, we apply our MB-B (best) on Rotten Tomatoes datasets and compare it with the state-of-the-art models, PlanSum and TransSum. As shown in Table 3, our approach achieves the best results, which shows that the results of our approach are reliable.

Model	**Rotten Tomatoes**		
	R-1	R-2	R-L
PlanSum	21.77	6.18	16.98
TransSum	25.34	8.62	18.35
MB-B	**26.00**	**9.07**	**18.92**

Table 3: The evaluation results on Rotten Tomatoes.