A Supplementary Material

A.1 Evaluation Classifiers

We train the style classifier to identify the styles on the target domain. The domain classifiers are trained to distinguish the samples from different domains. After training, all classifiers are used for evaluation only. The test accuracy of evaluation classifiers are reported in Table 8.

Style Classifier		Domain Classifier	
Dataset	Accuracy	Dataset	Accuracy
Yelp	97.6%	IMDB & Yelp	94.8%
Amazon	81.0%	IMDB & Amazon	97.1%
Yahoo	99.4%	IMDB & Yahoo	86.9%
ENRON	87.0%	GYAFC & ENRON	89.7%

Table 8: Test accuracy of evaluation classifiers.

A.2 Source Domain Data

To investigate the effectiveness of the source domain data, we evaluate our proposed models on different source domains that have unknown styles or the same styles as Yelp. Results are included in Table 9. It can be seen that the proposed models can robustly achieve favorable style transfer with help of different source domain data. Since DAST-C model mainly learns the generic content information by modeling the large corpus on the source domain, the number of source training data significantly affects the performance, especially on content preservation (BLEU). On the other hand, since DAST also adapts generic style information, the source domain that has closer sentiment information (IMDB) provides more benefit to the target domain (Yelp) than the TripAdvisor dataset does.

Model	Source	# samples	D-acc	S-acc	BLEU
DAST-C	IMDB	572K	96.9	90.3	17.8
	Yahoo	900k	90.3	91.3	19.6
	GYAFC	206k	93.5	92.9	16.1
DAST	IMDB	334k	97.0	92.6	20.1
	TripAdvisor	572k	86.2	91.4	18.4

Table 9: Performance on the Yelp (1% data) dataset with help of different source domain data.

A.3 Human Evaluation

For each human evaluation on Yelp sentiment transfer and Enron formality transfer tasks, we randomly sampled 100 sentences from the corresponding test set and collected three responses for each pair on every evaluation aspect, yielding 2700 responses in total. Each pair of system outputs was randomly presented to 7 crowd-sourced judges, who indicated their preference for style control, content preservation and fluency using the form shown in Figure 3. To minimize the impact of spamming, we employed the top-ranked 30% of U.S. workers provided by the crowd-sourcing service. In order to make the task less abstract, following Mir et al. (2019), we asked the judges to evaluate the content preservation quality independently of style information. Detailed task descriptions and examples were also provided to guide the judges. Inter-rater agreement, as measured by agreement with the most common judgment was 75.9%.

Besides the style control, content preservation and fluency evaluated in Table 3, we also asked each worker to provide a judgment of **the overall quality** in terms of three aspects as a whole. Results are summarized in Table 10. It shows that our DAST model is better in the overall quality compared to the baselines.

Overall Quality (Yelp 1% data)				
Our Model		Neutral	Comparison	
DAST	**81.1%**	14.0%	4.9%	ControlGen
DAST	**31.4%**	43.0%	25.6%	DAST-C
DAST	16.9%	23.9%	**59.2%**	human
Overall Quality (Enron)				
Our Model		Neutral	Comparison	
DAST	**52.7%**	35.3%	12.0%	ControlGen
DAST	**34.0%**	48.4%	17.6%	DAST-C
DAST	12.0%	17.8%	**68.0%**	human

Table 10: Results of **Human Evaluation** in terms of the overall quality on Yelp sentiment transfer and Enron formality transfer tasks.

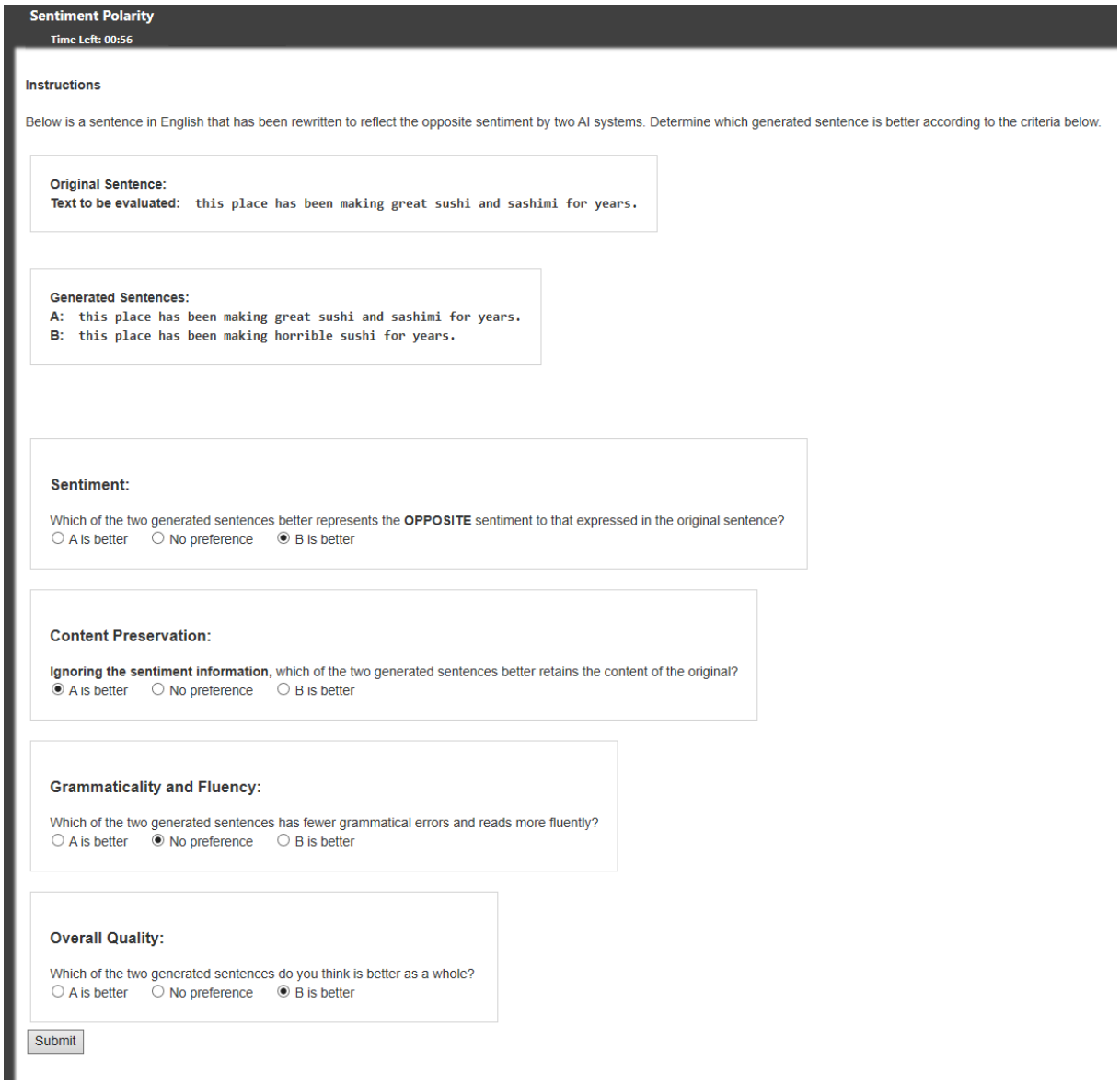

Figure 3: Questionnaire used to elicit pairwise judgments from crowd-sourced annotators. Candidate responses were presented in random order.