A Newsroom human evaluation prompt questions

Dim.	Question
IN	How well does the summary capture the key points of the article?
RL	Are the details provided by the summary consistent with details in the article?
VE	How efficient do you think the summary conveys the main point of the article?
UC	How much unnecessary content do you think the summary contains?
SR	To what degree do you think the summary is a perfect surrogate of the article?
CN	How much additional informative information can a reader find from the article after reading the summary?

Table 1: Prompts presented to Amazon Mechanical Turk workers

B Newsroom Pearson correlation results

ROUGE	**CN**	**IN**	**RL**	**SR**	**UC**	**VE**
R1-R	0.477	**0.957**	0.912	0.946	0.846	0.793
R1-P	0.570	0.770	0.841	0.836	0.886	**0.971**
R1-F	**0.662**	0.928	**0.963**	**0.964**	**0.967**	0.964
R2-R	0.301	0.876	0.823	0.891	0.787	0.790
R2-P	0.416	0.788	0.813	0.843	0.848	0.919
R2-F	0.444	0.848	0.859	0.898	0.881	0.923
Enc-ref	**CN**	**IN**	**RL**	**SR**	**UC**	**VE**
enc-2	0.688	0.987	0.990	**0.998**	0.974	0.889
enc-3	**0.714**	0.983	**0.994**	0.995	**0.981**	**0.896**
max	**0.714**	0.992	0.992	0.989	0.965	0.855
avg	0.697	**0.994**	0.991	0.994	0.965	0.860
InferSent	**0.714**	0.993	0.992	0.992	0.967	0.855
ELMo-m	0.702	0.993	0.990	0.991	0.965	0.852
ELMo-a	0.681	**0.994**	0.989	0.991	0.957	0.861
enc-avg	0.701	0.991	0.993	0.995	0.971	0.875
Enc-doc	**CN**	**IN**	**RL**	**SR**	**UC**	**VE**
enc-2	0.514	0.984	0.940	0.974	0.895	0.811
enc-3	0.556	0.990	0.956	0.982	0.914	**0.829**
max	0.677	0.997	0.973	0.980	0.934	0.788
avg	**0.734**	0.990	**0.985**	0.979	**0.953**	0.813
InferSent	0.679	0.996	0.971	0.977	0.927	0.781
ELMo-m	0.682	0.998	0.981	0.984	0.941	0.813
ELMo-a	0.687	0.993	0.976	0.972	0.923	0.795
enc-avg	0.624	**0.999**	0.972	**0.986**	0.929	0.817

Table 2: Pearson correlation on Newsroom human evaluation data. Enc-ref/doc refers to embedding similarity with reference/full document.