Investigating Crowdsourcing Protocols for Evaluating Factual Yale **Consistency of Summarization Xiangru Tang¹**, Alexander R. Fabbri¹, Ziming Mao¹, Griffin Adams³, Haoran Li², Yashar Mehdad², Dragomir Radev¹

Introduction

Large language models can create fluent summaries with high coverage, yet it is now well-documented that summaries often suffer from factual these inconsistencies. Though many automatic metrics have proposed to characterize the level of been hallucinations and factual errors, manual evaluation is recognized to be needed. Manual evaluation is essential to judge progress, yet not as much work has been done to address the ideal human evaluation setup. In this paper, we crowdsource annotations for factual consistency across Likert and Ranking-based evaluation scales across the CNN-Daily Mail and XSum datasets. We find that the BWS protocol produces more reliable annotations in general. But the Likert scale is not bad to be used for factuality evaluation in some cases, contrary to the argument in prior work. Based on our quantitative experiments, we also present the prerequisites of using Likert.

Study Design

Previous work has analyzed the biggest limitations of modern pre-trained models for abstractive document summarization and found that these models are highly prone to hallucinate content that is unfaithful to the input document. Thus, we conduct studies on the faithfulness task to elicit summary judgments for analysis. In the task, we ask annotators to judge the factual consistency of the summaries. We organize our findings along three main research questions (RQ) outlined in this section.

- RQ1: Ranking (BWS) vs. Likert on factual consistency.
- RQ2: Considering different models and datasets, how is the reliability of scale methods?
- RQ3: How serious is the four limitations we defined of these scale methods?
 - 1. Inconsistencies in annotations by the same annotator: an annotator might assign different scores to the same item when the annotations are spread over time.
 - 2. Inconsistencies in annotations by different annotators: one annotator might assign a score of 7 to the word good on a 1-to-9 sentiment scale, while another annotator can assign a score of 8 to the same word.

Table 1. Overall results of baselines across the tasks of XSum.

Model	Rouge-1	Rouge-2	Rou
PEGASUS	46.84 ¹	24.52 ¹	39.1
ProphetNet	43.23 ³	19.96 ³	35.1
BART	44.15 ²	21.28^{2}	35.9
BERTSUM	38.21 ⁴	16.11 ⁴	30.8

 Table 2. Overall results of baselines across the
 tasks of CNN/DM.

Model	Rouge-1	Rouge-2	Ro
PEGASUS	44.19 ¹	21.45 ¹	41.
ProphetNet	42.45^{3}	19.90 ³	39.
BART	44.07^{2}	21.13 ²	40.
BERTSUM	41.82^{4}	19.39 ⁴	38.

Table 3. The average Likert scores and the average rankfor
 all systems on XSum and CNN/DM.

Madala		XSum	8	CNN	I/DN
Models	BWS	Likert	Likert-10	BWS	Likert
PEGASUS	3.247^{3}	3.350^{1}	6.247^{2}	3.230^2	3.887^{2}
ProphetNet	3.360^2	3.293 ³	6.427^{2}	3.100^{3}	3.860^{4}
BART	3.570 ¹	3.433 ²	6.937 ¹	3.593 ¹	4.017 ¹
BERTSUM	2.827^4	2.790^{4}	5.163^{4}	3.087^4	3.863 ³

Table 4. Instance-level reliability computed by Krip-pendorff's alpha (α) on the CNN/DM and XSum

krippendorff	XSum	CNN/DM
BWS	0.2477	0.1582
Likert	0.2951	0.0443

Table 5. System-level split-half reliability computedby Split-Half Reliability (SHR) on the CNN/DM andXSum

	SHR	XSum	CNN/DM
	BWS	0.90314	0.87657
23	Likert	0.92777	0.45619

Yale University¹ Facebook Al² Columbia University³

ıge-L 10¹ 83⁴

ouge-L .08¹ .31³ $.89^{2}$ $.67^{4}$

Figure 1. Score distribution of Likert for faithfulness. Each data point shows the number of times a particular score was assigned to each system



Figure 2. Score distribution of Likert scale 1-10 for faithfulness. Each data point shows the number of times a particular score was assigned to each system



Figure 3. Mturk interface

Your task						-	-	
Your task is to evalua summaries that repre	te the machine sent the most	e-generated te important inf	ext summaries ormation of th	. Text summa	rization is the	process of c	reatin	g
In each of the E costi	and you will be	o proported w	ith 4 different	cummorios ex		a an article		
in each of the 5 secto	ons, you will be	e presented w	iur 4 amereni	summaries co	prresponding o	o an article.		
After you finish evalu	ating the 20 su	ummaries in 5	sections, you	will receive an	n acknowledge	ment code.		
Fill out that code on t	he MTurk page	Ð.						- 1
Turker ID								
Short answer text								
r section 1 Continue	e to next sectio	on –						
ection 2 of 7 Instructio Rate the summaries I	e to next section	on _	stency with th	e source. Fact	ual consistenc	y is rated on	aten	÷
ection 2 of 7 Instructio Rate the summaries I point scale where 10	e to next section DS pased on their means perfect	factual consi	stency with th	e source. Fact means very po	ual consistenc	y is rated on sistency.	a ten	ž.
r section 1 Continue ection 2 of 7 Instructio Rate the summaries t point scale where 10 The factual consister Factual consistercy r summary will certain	next section TS based on their means perfect icy of a summing ay not always y be a bad sun	factual consis factual consis factual consi ary is determi s relate to how nmary.	stency with the istency and 1 ned by its agre v good the sur	e source. Fact means very pr eement with fa nmary is, thou	ual consistenc or factual con acts in the sou ggh a factually	y is rated on sistency. rce documen inconsistent	a ten	
ection 2 of 7 Instruction Rate the summaries I point scale where 10 The factual consister Factual consistercy r summary will certain In some cases, you m of machine-generated number.	e to next section Dased on their means perfect hey of a summ. nay not always y be a bad sur lay find that th d summaries. F	factual consis factual consis t factual consi ary is determi s relate to hov mmary. e article and t Please indicat	stency with th istency and 1 ned by its agree v good the sur the summaries is so at the eni	e source. Fact means very pr eement with fa nmary is, thou s do not match d of the form to	ual consistenc or factual con acts in the sou gh a factually o, this may be (with the sectio	y is rated on sistency. rcc documer inconsistent due to the loo n and the su	a ten nt. mmar	: lity y
ection 2 of 7 Instruction Rate the summaries to point scale where 10 The factual consistency of summary will certaind In some cases, you m of machine-generated number. If you have any quest	e to next section Dased on their means perfect icy of a summ. hay not always y be a bad sum hay find that th d summaries. Fi ions or there is	factual consis factual consis t factual consi ary is determi are late to hov mmary. e article and t Please indicat s an issue with	stency with the istency and 1 in ned by its agree y good the sur the summaries the summaries the summaries the summaries the summaries	e source. Fact means very po sement with fa mmary is, thou ammary is, thou d of the form n ase email lily!a	ual consistenc or factual con cts in the sou cgh a factually n, this may be (with the sectio abmturk@gma	y is rated on sistency. rcc documer inconsistent due to the lo' n and the su il.com	a ten nt. mmar	i . lity y
r section 1 Continue ection 2 of 7 Instructio Rate the summaries I point scale where 10 The factual consistency r Factual consistenc	e to next section DS based on their means perfect tory of a summi- nay not always be a bad sum- nay not always be a bad sum- summaries. Fi ions or there is independent of the summaries of the independent of the summaries of the summaries of the summaries of the independent of the summaries of the summaries of the summaries of the independent of the summaries of the summaries of the summaries of the independent of the summaries of the summar	factual consis factual consis t factual consi ary is determi s relate to hov mmary. e article and t Please indicat s an issue with e instruction	stency with the istency and 1 ined by its agra- v good the summaries the summaries is a at the ener- the data, plei- ins? *	e source. Fact means very pr eement with fa nmary is, thou s do not match d of the form n ase email lilyla	ual consistency or factual con acts in the sou gh a factually o, this may be o with the sectio abmturk@gma	y is rated on sistency. roe documer inconsistent due to the lo n and the su il.com	a ten nt. mmar	: Hity Y
r section 1 Continue ection 2 of 7 Instructio Rate the summaries t point scale where 10 The factual consistency 7 summary will centiani In some cases, you m of machine-generated number. If you have any quest How well do you un	e to next section TSS based on their means perfect toy of a summ. nay not always y be a bad sun ay find that the d summaries. F ions or there is inderstand the 1	factual consist factual consist factual consi ary is determi s relate to hov mmary. e article and t Please indicat s an issue with e instruction 2	stency with the istency and 1 in ned by its agrin v good the sur the summaries is so at the en- the data, plei- ns? * 3	e source. Fact means very po eement with fa mmary is, thou d of the form of ase email lilylo	ual consistenc or factual con acts in the sou gh a factually n, this may be to with the sectio abmturk@gma	y is rated on sistency. rcc documer inconsistent due to the lo n and the su il.com	a ten nt. mmar	i Hity y



Results

We use both Likert and BWS on the output of four recent summarizers. The majority of evaluations is conducted using judgements, with the second most frequent method being rank-based annotations. Best-worst scaling (BWS) is a specific type of ranking-oriented evaluation that requires annotators to specify only the first and last rank, which is claimed to be reliable. We are using BART, ProphetNet, Pegasus, and BERTSUM models on XSUM and CNN/DM. Each study contains 100 original passages, and 400 summaries generated from four pre-trained models. We randomly sample 100 documents from the corpus with corresponding summaries from models to form the item set for all our studies. We demand 3 judgments per summary to mirror a common setup in manual evaluation, while each annotator only can annotate 5 summaries. We separated our corpus into 20 blocks of 5 documents and included all 4 generated summaries for each document in the same block, which results in 5 \times 4 = 20 summaries per block.

Conclusion

- and cost-efficient than Likert scale.
- increases the noise and bias.

Acknowledgement

Thanks to Drago and my collaborators for this project. Thanks, Alex, Ziming. This couldn't happen without their insights, efforts, and leadership.



LILY Lab

3. Scale region bias: annotators often have a bias towards a part of the scale, for example, preference for the middle of the scale.

4. Fixed granularity: in some cases, annotators might feel too restricted with a given rating scale and may want to place an item inbetween the two points on the scale. On the other hand, a fine-grained scale may overwhelm the respondents and lead to even more inconsistencies in annotation.

• Regrading RQ1, on the general used CNN/DM, we show that ranking-style evaluations (BWS) are more reliable

• Regrading RQ2, We found different phenomena happened on different datasets XSum, while BWS and Likert are comparable, Likert even is a little better. Because in CNN/DM, it's easy to distinguish which one is better (we see many 4 and 5 in Figure 1), which benefits BWS. But in XSum, swinging between close scores