Dimensions	**Description**	**What you fill in**	**Example**	**Comments**
Key	Paper ID, maps to here: https://drive.google.com/drive/u/2/folders/15-4WoUw0E0JNRlOS	n/a		
Paper	Paper title	n/a		
Conf	conference acronym	conference acronym	EMNLP	
Year	year	year	2018	
Task(s)	What ST tasks are covered in paper?		formality, sentiment	
Human Annotation?	Is human annotation present? ; if No, then no more work for this paper	Yes I No	Yes	
Who are the annotators?	Who did the annotation: AMT or Figure 8, etc.	<free text>	11 CL grad students who are native English spe	experts, non-experts, previously known to the authors
Annotators payment?	How much are annotators compensated?	<free text>	$0.3/HIT	unspecified
Quality control	Was there QC done on the annotations?	Yes I No	qualification task	either in the form of spotchecking (manual, automatic), throwing out the judgments of low performing raters, or pre-task qualification tasks
Availability of annotations	Are the human judgments downloadable from some site?	Yes I No	No	
# Systems	How many systems are evaluated? (this includes reference)	<number>	5	
# Sent/system	How many instances per system are evaluated?	<number>	100	
Annotators per sentence	How many annotations per instance are collected?	<number>	3	
IAA	Was annotator agreement calculated?	Yes I No	No	
Sampling method	How are system outputs (or other evaluation items) selected for inclusion in the evaluatic	<free text>	random	
Other Comments	Any other information relevant to the overall task design can be added here.	<free text>	this paper is an arxiv version	
Present?	Do they perform human evaluation in this dimension (style)?	Yes I No	Yes	
Quality Criterion	What is the quality criterion name (as mentioned in the paper)?	<free text>	formality	
Absolute judgment	Is each output assessed for quality in its own right **(as opposed to other system outputs**	Yes I No	No	
Relative judgment type (if applicable)	What is the type of relative judgment ? (see comments for more details)	pairwise I ranking I best selection	pairwise	**This option is only applicably if the answer to Question 21 is No.** **pairwise:** given exactly 2 system outputs choose the best **ranking:** given more than 2 system outputs rank all of them **best selection:** given more than 2 system outputs choose the best
Absolute rating scale (if applicable)	What is the scale of the rating (i.e. what are the different possible values annotators can c	list	[-3,-2, -1, 0, 1, 2, 3]	**This option is only applicably if the answer to Question 21 is Yes.**
Lineage / Duplication	Does the paper include reference to instructions/annotation scheme of previous paper?	Yes I No	Yes	
Lineage Source (if applicable)	If they leverage prior work, what is the cite for the work?	cite of the paper, or acronym	GYAFC, Shen et al. (2017)	**This option is only applicably if the answer to Question 25 is Yes.**
Other	any other information relevant to the overall task design can be added here	<free text>		
Present?	do they perform human evaluation in this dimension (meaning)?			
Quality Criterion	for example sometimes people say "meaning preservation" or "content preservation"			
Absolute judgment				
Relative judgment type (if applicable)				
Scale of rating **absolute** judgment (if applicable)				
Lineage / Duplication				
Lineage Source (if applicable)				
Other				
Present?	do they perform human evaluation in this dimension (fluency)?			
Quality Criterion	for example sometimes people say fluency or grammaticality			
Absolute judgment				
Relative judgment type (if applicable)				
Scale of rating absolute judgment (if applicable)				
Lineage / Duplication				
Lineage Source (if applicable)				
Other				
Present?	do they perform human evaluation in this dimension (other / overall)?			
Quality Criterion				
Absolute judgment				
Relative judgment type (if applicable)				
Scale of rating absolute judgment (if applicable)				
Lineage / Duplication				
Lineage Source (if applicable)				
Other				
Other Notes				
if you can't find the answer in the paper, write "not available"				

Dimensions	Description	What you fill in	Example	Comments
if you a question is not applicable write "n/a" **as opposed to leaving the cell blank!** (that way we can distinguish between missing annotations and n/a), e.g, if a paper does not perform human evaluation at all most of its cells should be annotated as "n/a"				
if you are not sure about a response but a question mark (?) after it.				
if references are part of the evaluation count this as a separate system				
if some of the information in columns 2-16 is	100 (meaning); 60 (fluency)			
if you are highlhy uncertain about some of your annotations and want a second person to look at it, leave this as a comment				