EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

SFB 833
Projekt A4

Kordula De Kuthy, Ramon Ziai & Detmar Meurers

Focus Annotation of Task-based Data: Establishing the Quality of Crowd Annotation

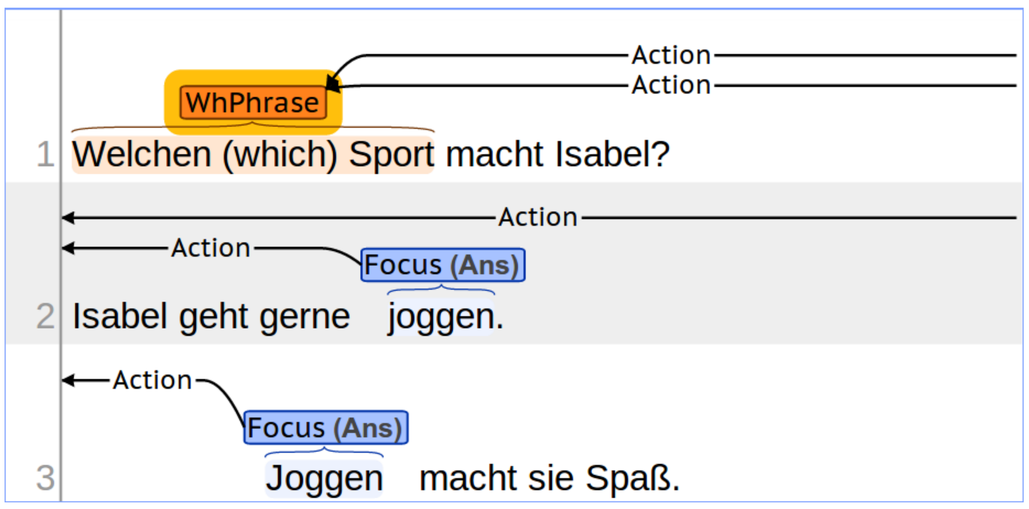
Motivation

- Inter-annotator agreement studies for focus annotation found it is difficult to obtain high levels of agreement.
- Reliable expert focus annotation is feasible if one provides an task context with explicit questions (Ziai & Meurers, 2014).
- Can "the crowd" provide reliable focus annotation?
- When and where is the crowd annotation reliable?
  - Hypothesis: Sentences with little variation in the annotation are more reliably annotated, i.e., are of a higher quality.

Authentic data: CREG (Ott, Ziai & Meurers, 2012)

- German reading comprehension corpus
- We used the CREG-5K subset, which contains:
  - 877 reading comprehension questions
  - 966 target answers by the teachers
  - 5,138 learner answers
  - avg. length of learner answer: 11.58 tokens
- Each student answer was rated by two annotators with respect to whether it properly answers the question.

Expert Focus Annotation (Ziai & Meurers, 2014)



- Our annotation scheme is incremental and consists of
  - **Question Form**: surface form of question
  - **Answer Type**: semantic category of focus in relation to question
  - **Focus**: focused words or phrases in answer
- Agreement (token level, macro-averaged by answer):

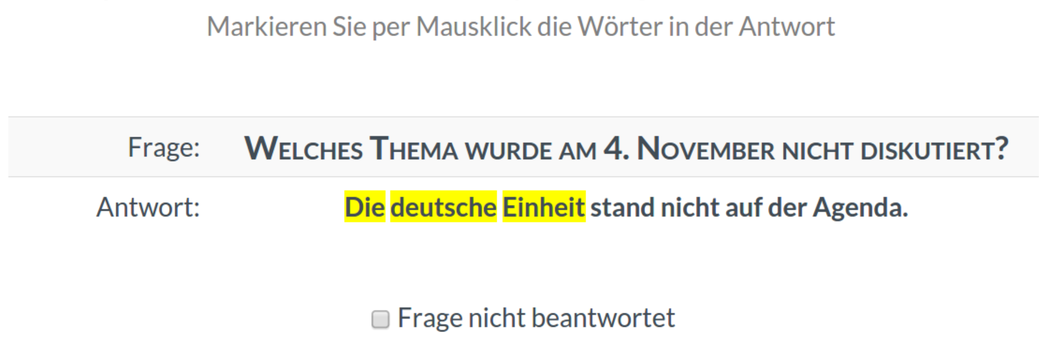
	% Agreement	κ
CREG-1032	88.1%	0.75
CREG-2155	86.3%	0.70
Overall	86.6%	0.71

$\Rightarrow$ substantial agreement

- *Gold standard annotation* obtained by merging the two annotations, with a judge deciding in cases of conflict.

Crowd Focus Annotation (De Kuthy, Ziai & Meurers, 2015)

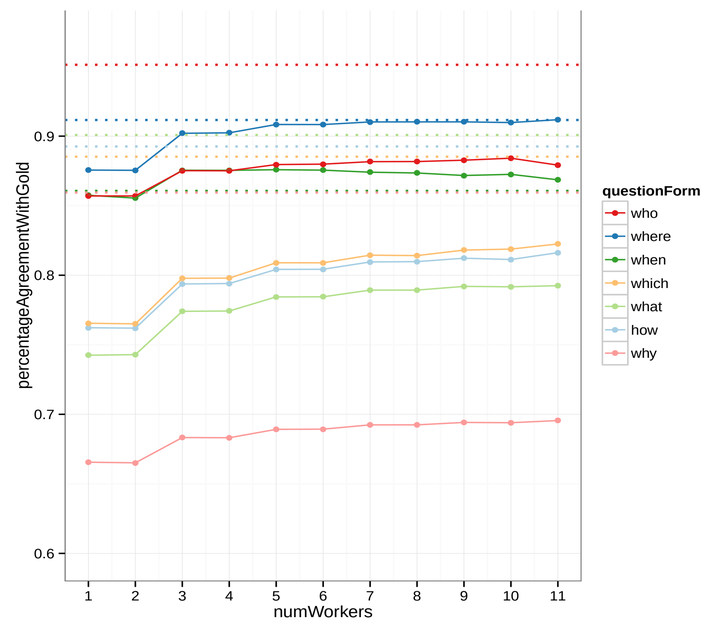
- Setup of Annotation Study using Crowdflower:



- Comparing crowd and expert annotation:
  - For each number of workers (1, . . . , 11), determine all possible combinations of worker judgments.
  - Calculate the combined judgments for each word by majority voting, breaking ties by random assignment.
  - Use percentage agreement to compare crowd and experts.

Quality: Crowd Agreement with Gold Standard

- More workers provide more robust judgements
- *who*, *when*, and *where* questions support reliable focus annotation
- *which*, *what* and *how* questions more difficult
- *why* is most difficult for the crowd, while experts do far better
  $\rightarrow$ guidelines important



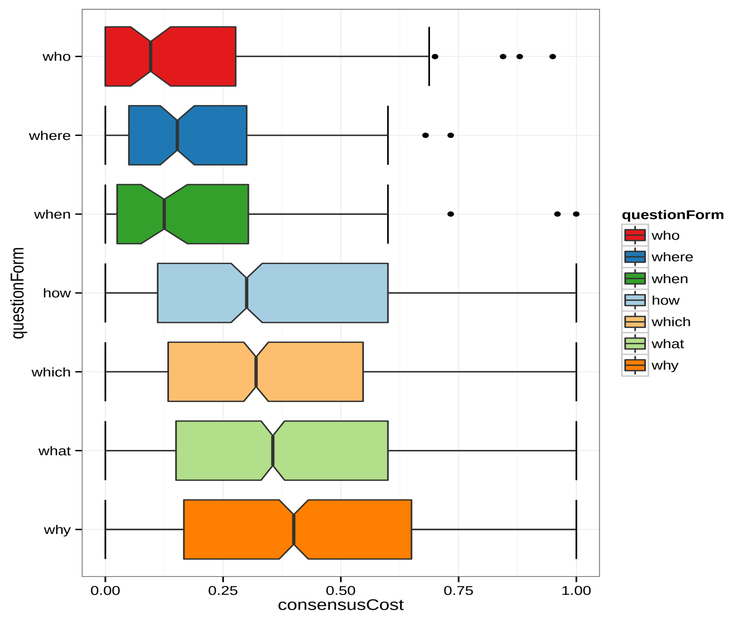
Quality: Agreement within the Crowd

- Measure the diversity of the focus annotation in terms of the *Consensus Cost (CC)*:

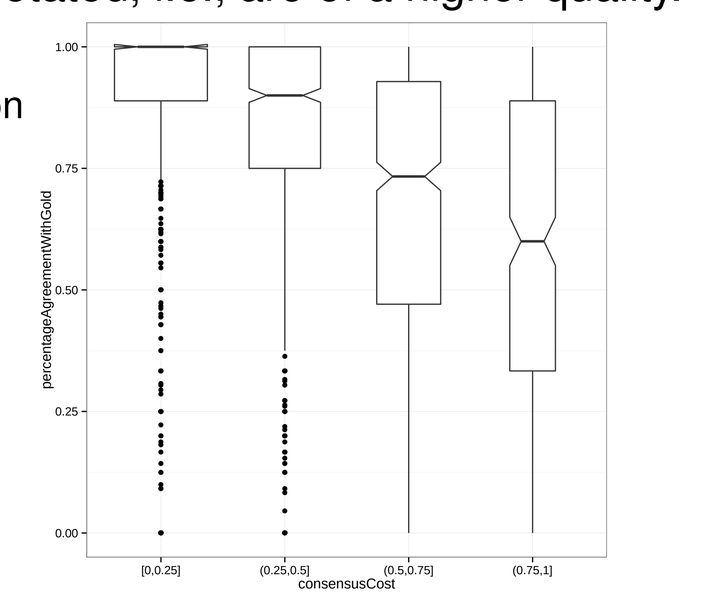
$$CC = \frac{\sum_{w=0}^{n} changeNeededForConsensus(w)}{maxChangesNeeded \times n}$$

- Analysis by question form:

$\Rightarrow$ Consensus cost by question from patterns parallel to the quality of the crowd annotation.



- Sentences with little variation in the crowd annotation are more reliably annotated, i.e., are of a higher quality.
- Confirmation of hypothesis by comparison with gold standard:
  - Low consensus cost (< 0.5)
    $\Rightarrow$ high quality annotation
  - High consensus cost
    $\Rightarrow$ heterogeneous quality



Extrinsic Evaluation

- Establish relevance and quality of the focus annotation
- We used the short answer assessment system CoMiC
  - analyzes the quantity and quality of alignments between student and target answer
- Exploring the impact of different Consensus Costs:
  - Four cutoffs: 0.25, 0.5, 0.75, 1.0.
  - use the answers with crowd focus annotations satisfying the cutoff constraint in training and test set

Cost $\leq$	Focus train/test	%	Given train/test	%	Avg %
base	–		4136/1001	81.5	81.5
0.25	1009/252	88.1	3127/749	80.4	82.3
0.5	2019/489	84.5	2117/512	80.7	82.5
0.75	3087/747	84.5	1049/254	79.5	**83.2**
1.0	3638/882	82.7	498/119	76.5	81.9

For more information, check out http://purl.org/icall/comic