Indirect question answering in English, German and Bavarian

A challenging task for high- and low-resource languages alike

Miriam Winkler, Verena Blaschke, Barbara Plank

New trilingual IQA datasets

English, German, Bavarian
(non-std dialect related to German)

InQA+

- **Evaluation data** (400+ instances)
- **Manually translated & annotated**
- **Parallel data**

⚠️ **Disagreement among annotators**
(especially cases that are not straightforwardly yes/no)
– common for IQA datasets!

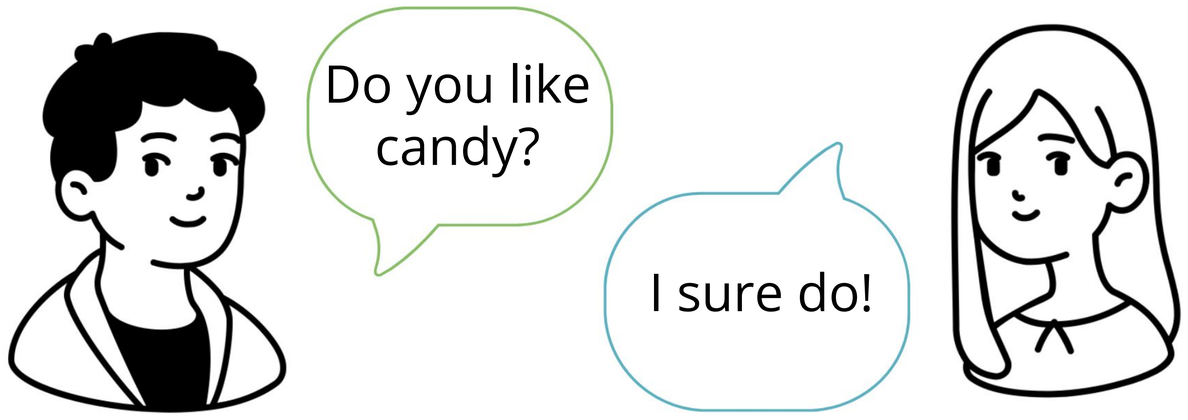
GenIQA

- **Training data** (1.5k instances)
- **LLM-generated** (GPT-4o-mini)

⚠️ **Generated dialogue/label combinations can be unintuitive**

⚠️ **LLMs struggle with generating Bavarian:** generated text is often in German, code-switched, or contains linguistic mistakes

Indirect QA



Experiments

mBERT, mDeBERTa, XLM-R

- All classification models perform poorly on all languages

Quality vs. quantity of training data

- Small/manual (615 English instances via *IndirectQA*): higher accuracy
- Large/generated (GenIQA): higher macro F1
- Highest accuracy/F1 when adding lots of training data from other manually annotated dataset (6k instances)

Conclusions

- **Data quantity is important,** but cannot trivially be solved through automatically generating training samples
- Data for pragmatically challenging tasks like indirect QA are **hard to annotate**
  – can we improve agreement / find causes of disagreement?

	InQA+	GenIQA		
Yes	183	400	250	218
No	100	132	210	182
Cond. yes	18	319	274	306
Neither/nor	79	477	464	468
Other	34	71	141	191
Lacks context	24	101	161	135
Total	3×438	1500	1500	1500