Yes, No or IDK: The Challenge of Unanswerable Yes/No Questions

Elior Sulem, Jamaal Hay and Dan Roth

Department of Computer and Information Science
University of Pennsylvania

NAACL 2022

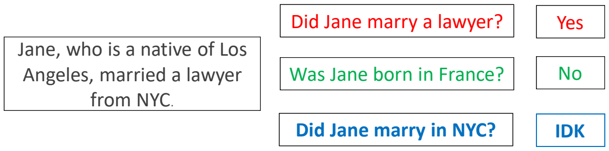
Yes/No Question Answering

- **Task:** Given a paragraph and a naturally occurred question, answer Yes **or** No**.**
- **Dataset: BoolQ (Clark et al., 2019)**

Context1: Unlike trophies such as the Stanley Cup and the Grey Cup, a new Vince Lombardi Trophy is made every year and the winning team maintains permanent possession of that trophy, with one notable exception being Super Bowl V's, won by the then-Baltimore Colts. The city of Baltimore retained that trophy as part of the legal settlement between the team and the city after the Colts' infamous ``Midnight Mayflower'' move to Indianapolis on March 29, 1984. Since then, both the relocated Colts and their replace in Baltimore, the Ravens, have won the Super Bowl and earned trophies in their own right.
Question1: Do they make a new Lombardi Trophy every year?
Answer: Yes

Example of IDK example from BoolQ (Clark et al., 2019) whose answer is Yes.

Yes/No Questions can be Unanswerable


Jane, who is a native of Los Angeles, married a lawyer from NYC.

Did Jane marry a lawyer? Yes

Was Jane born in France? No

Did Jane marry in NYC? IDK

- In practical situations, we do not always have the information required to answer the question.
- In these cases, we expect the system to answer "IDK".

Augmenting BoolQ with IDK Questions

- Sampling randomly half of the "Yes" questions and half of the "No" questions
- **Matching to each of the extracted questions a passage from BoolQ that has the greatest overlapping with the questions in terms of nouns and verbs.**
- We apply this process on both training and dev sets.
- Similar method was used in Clark and Gardner (2018) in the case of Extractive QA.

Context2: The Vince Lombardi Trophy is the trophy awarded each year to the winning team of the National Football League's championship game, the Super Bowl. The trophy is named in honor of NFL coach Vince Lombardi.
Question1: Do they make a new Lombardi Trophy every year?
Answer: IDK

Example of IDK example obtained by swapping existing contexts and questions. Content words that appear in both the context and the question are underlined.

Architecture used: BERT-LARGE (Devlin et al., 2019)
Existing datasets used: BoolQ , SQuAD 2.0 (Rajpurkar et al., 2018), MNLI (Williams et al., 2018)

BoolQ$_{3L}$ Dataset

New Dataset

We add to BoolQ (Clark et al., 2019) the obtained IDK questions.
Validation: 95% correct in dev (sample of 100 examples), absolute IAA=100%; **93.5% correct in train** (sample of 100 examples, absolute IAA= 97%.

Portions	#Examples	IDK Prportion (%)
Train	14.1K	33
Dev	4.9 K	33

Statistics for the BooQ$_{3L}$ dataset

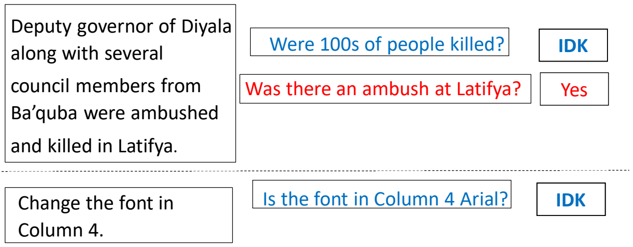
Out-Of-Domain Test Sets

New Test Datasets

- **ACE-YNQA: Event-related Yes/No Questions with IDK option**, semi-automatically derived from the ACE event extraction dataset (Walker et al., 2006), focusing on **time and location arguments**.
- **INSTRUCTIONS**: **Instruction-related Yes/No Questions with IDK option**, manually derived from scratch.

Portions	#Examples	IDK Prportion (%)
ACE-YNQA	999	52
INSTRUCTIONS	70	33

Statistics for the Out-Of-Domain (OOD) Test Sets


Deputy governor of Diyala along with several council members from Ba'quba were ambushed and killed in Latifya.

Were 100s of people killed? **IDK**

Was there an ambush at Latifya? Yes

Change the font in Column 4.

Is the font in Column 4 Arial? **IDK**

Examples from ACE-YNQA (top) and INSTRUCTIONS (bottom)

Yes/No QA: 3 Labels vs. 2 Labels

	Baseline	Using RTE	Using Binary RTE
train → ↓ test	**BoolQ$_{3L}$**	**MNLI + BoolQ$_{3L}$**	**c(MNLI) +BoolQ$_{3L}$**
BoolQ$_{3L}$ dev	33.64	42.66	**43.25**
ACE-YNQA	52.02	52.02	**54.94**

Accuracy of the BERT-LARGE system evaluated on BoolQ$_{3L}$ and ACE-YNQA (**3 Labels**)

train → ↓ test	**BoolQ**	**MNLI + BoolQ**	**c(MNLI) +BoolQ**
BoolQ dev	72.88	78.24	**79.49**
ACE-YNQA$_{Y/N}$	59.53	65.47	**68.01**

Accuracy of the BERT-LARGE system evaluated on BoolQ and ACE-YNQA$_{Y/N}$ (**2 Labels**)

Conclusion

- The ability to answer IDK is necessary for Yes/No QA in realistic scenarios.
- We extend BoolQ with unanswerable questions and provide OOD test sets.
- We show the difficulty of the 3-label task, compared to the 2-label task.