Entity-Based Knowledge Conflicts in Question Answering

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, Sameer Singh
Empirical Methods in Natural Language Processing (EMNLP) 2021 · Apple Inc. and UC Irvine

What Is This Work About?

We study QA models when faced with **knowledge conflicts (KCs)** - situations where contextual knowledge **contradicts** parametric knowledge.

Context: The **Chicago Bulls** defeated the **Seattle SuperSonics** in the 1996 NBA finals.

Q: Who won the 1996 NBA finals?
Ans: Chicago Bulls
Pred: Chicago Bulls

Context: The **Seattle SuperSonics** defeated the **Chicago Bulls** in the 1996 NBA finals.

Q: Who won the 1996 NBA finals?
Ans: Seattle SuperSonics
Pred: Chicago Bulls

QA models can memorize and regurgitate answers despite not having contextual support.

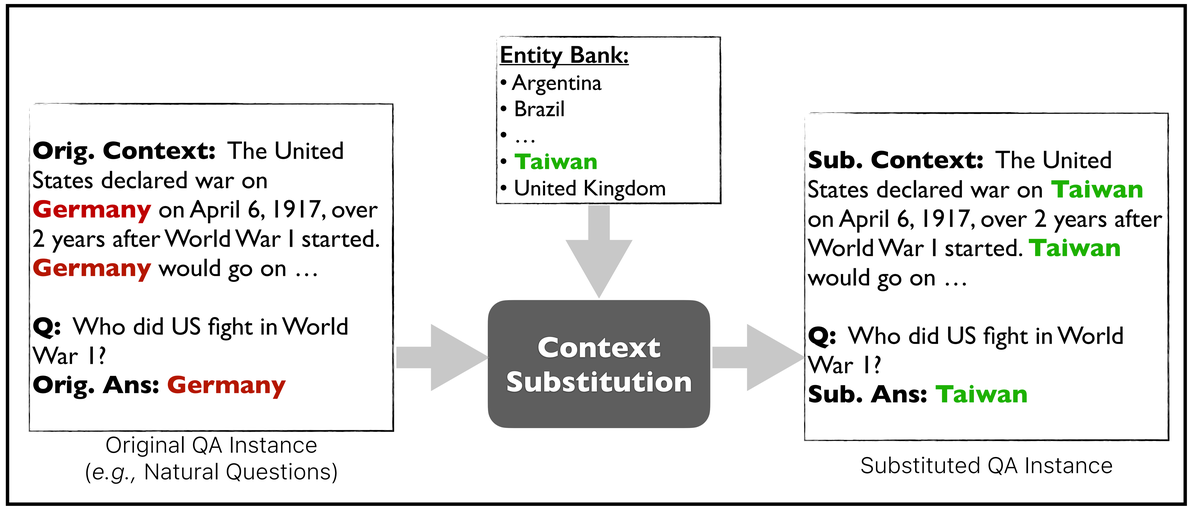
Why Care If Models Ignore Context?

Models which ignore contextual information and over-rely on parametric knowledge will:

- Fail to generalize to evolving temporal knowledge.
- Be less interpretable.
- Be prone to hallucinations, biased generations, and stochastic parroting.

Models which ground predictions in the context can mitigate these issues.

How Do We Create QA Instances To Test Knowledge Conflicts?



Original QA Instance (*e.g.*, Natural Questions)

Substituted QA Instance

Our substitution framework for generating substituted instances which contains information that contradicts what may have been learned during pre-training or fine-tuning, introducing a knowledge conflict.

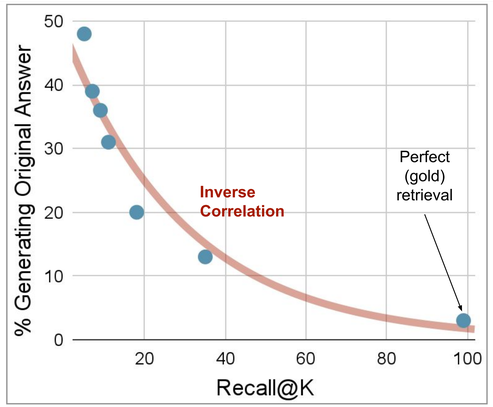
What Do Knowledge Conflicts Teach Us?

Primarily evaluate retrieve-and-read T5 QA models on our substituted contexts. **We find:**

- **Models over-rely on parametric knowledge**, regularly regurgitating the original answer despite context supporting a new answer (29.5% of time on Natural Questions).

Factors that contribute to hallucination:

- ⬆️ Model size → ⬆️ Hallucinate orig. answer.
- ⬇️ Retriever quality→ ⬆️ Hallucinate orig. answer.
- Orig. answer in train set → ⬆️ Hallucinate orig. answer.



The worse the retriever is when training, the more likely we will see models memorize instead of read.

Can We Mitigate Hallucinations?

Solution: Train on original QA instances **AND** on substituted QA instances from our framework.

Results:

- Hallucination of orig. answer drops significantly (29.5% → 2.6% on Natural Questions).
- Model learns to read and generalize better to OOD QA datasets (+4.4% EM on NewsQA).

What Else Can You Learn From Our Paper?

- How prevalent is memorization across different QA datasets?
- Does the popularity of an entity or the entity type affect how likely it is to be memorized?
- Do span-selection models suffer from an over-reliance on parametric knowledge?
- **How you can use our substitution framework to create custom substitutions and test your own hypothesis about model behavior!**

What Do We Release to the Community?

We have open-sourced the substitution framework we created to test knowledge conflicts:

github.com/apple/ml-knowledge-conflicts