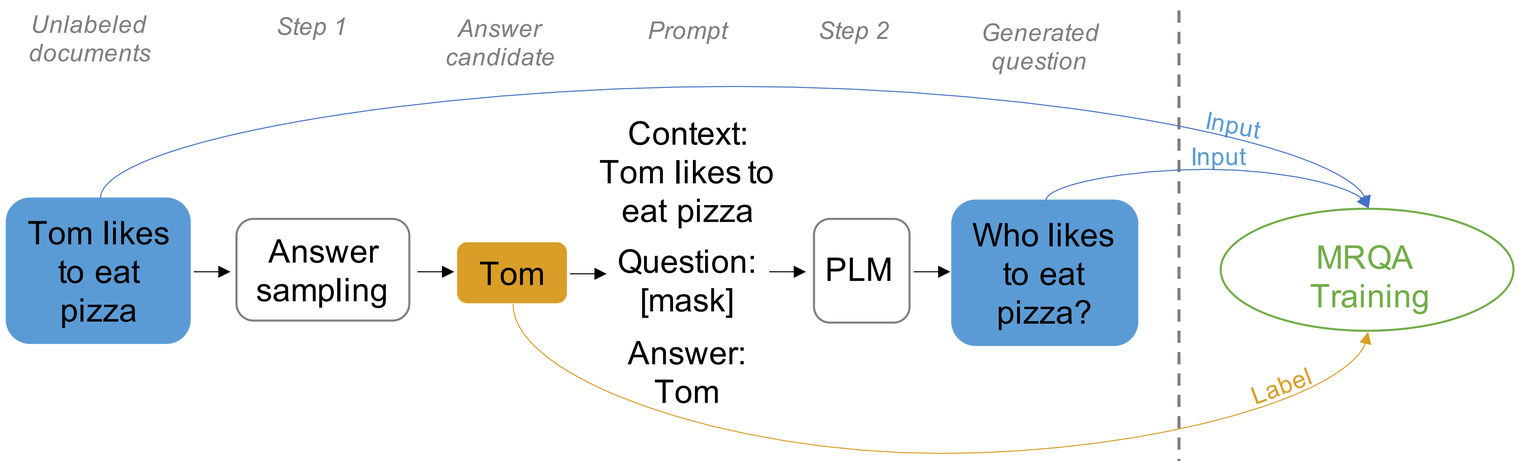
Prompting-based Synthetic Data Generation for Few-Shot Question Answering

Maximilian Schmidt, Andrea Bartezzaghi, Ngoc Thang Vu

LREC-COLING 2024 TORINO, ITALIA



Introduction

- Data generation has been shown to improve low-resource extractive Question Answering (QA) [4]
- LLMs encode linguistic knowledge, how can we make use of it for data generation?

⇒ Add task- and domain-specific unsupervised pre-training stage using data generation: Generate question-answer pairs from documents by Prompting LLMs

Method

Answer Sampling

- Sample answers using NER: Given context c this yields textual answer candidates a_c with spans
- NER is domain-agnostic and does not necessarily need data

Question Generation

- Prompt LLM to generate question q given c and a_c:

$$p(q|c, a_c) = \sum_{t=1}^{T} \log p(q_t|q_{<t}, c, a_c)$$

- We use T5 v1.1 large [2] with soft prompt:
  context: <context> question: <mask> answer: <answer>.
- Use provided documents or crawl from available sources, e.g., abstracts from PubMed
- Filter synthetic data using heuristics and consistency filtering for training QA model (which is also a Prompting (again using T5 v1.1 large) model in our case)

Code

Paper

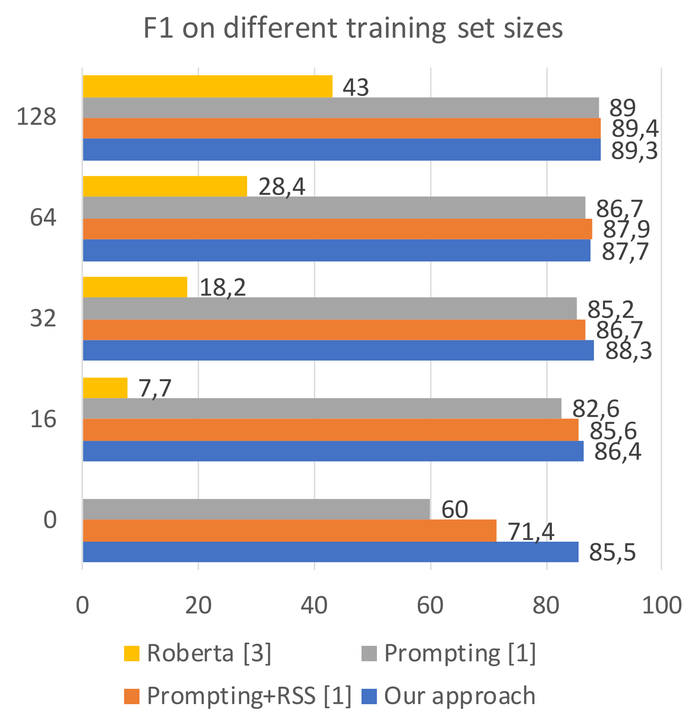
Experiments

Setup

- Few-shot MRQA benchmark [3]: 0, 16, 32, 64, 128 samples
- separate validation set of size 2048 using SQuAD for hyperparameter tuning

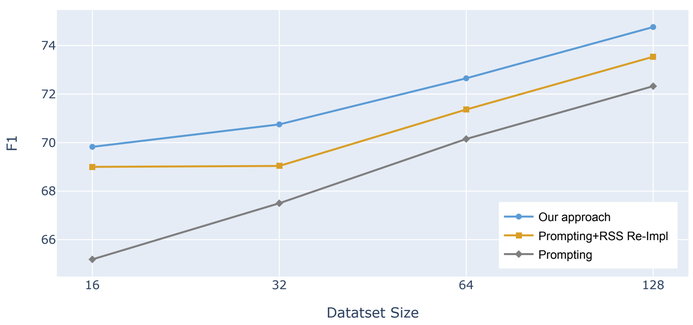
Results

1. How does Prompting-based data augmentation help in the zero- and few-shot setting?



Prompting-based data generation outperforms other approaches on SQuAD, especially in the zero-shot setting.

2. How does this approach generalize to other domains?



Our approach (with the hyperparameters from SQuAD) also improves the performance on all other domains in average.

3. How does NER-based answer sampling affect the performance?

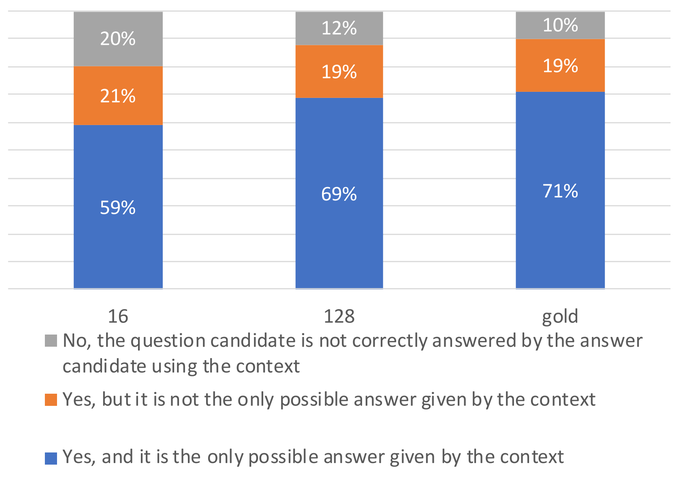
Model	0	16	32	64	128
sampled answer	85.5	86.4	88.3	87.7	89.3
gold answers	87.3	87.6	89.5	88.3	90.3
synthetic data only	87.3	90.0	91.0	90.7	91.3

Table: F1 of QA model using generated data from our approach on SQuAD in the zero- and few-shot setting (16, 32, 64 & 128 samples) as well as comparing gold answers with sampled answers.

NER-based answer sampling can be well suited for data generation for QA.

Analysis

- Analyze quality of generated data on NewsQA (worst performance)
- Run human study in which participants were asked to answer "Is the question candidate correctly answered by the answer candidate?"



More gold-labeled data improves quality of generated data, which is on par with gold-labeled data by taking into account only 128 gold-labeled samples

Conclusion

- Prompting-based data augmentation improves QA performance by exploiting LLM's linguistic knowledge, suggesting that LLMs are powerful when used for task- and domain-specific data augmentation
- Our approach is particularly suitable for the zero-shot setting

References

[1] Or Castel et al. *How Optimal is Greedy Decoding for Extractive Question Answering?* Nov. 2022.

[2] Colin Raffel et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* July 2020.

[3] Ori Ram et al. *Few-Shot Question Answering by Pretraining Span Selection.* June 2021.

[4] Siamak Shakeri et al. "End-to-End Synthetic Data Generation for Domain Adaptation of Question Answering Systems". In: (Oct. 2020).

Institut für Maschinelle Sprachverarbeitung

University of Stuttgart
Institute for Natural Language Processing

Code Repository:
`https://github.com/mxschmdt/mrqa-prompting-gen`