Stanford | ENGINEERING
Computer Science

Soft Contextual Data Augmentation for Out-of-Domain QA

Kevin Tien, Megumi Sano, Toby Frager (CS 224N 2022, Default Project, RobustQA Track)

Question answering using SOTA Transformer models shows brittleness to **domain transfer**. **Soft contextual data augmentation (SCA)** is a data augmentation strategy that has shown promising results on neural machine translation [1] and we apply this to out-of-domain QA.

Does soft contextual data augmentation improve performance on out-of-domain QA?

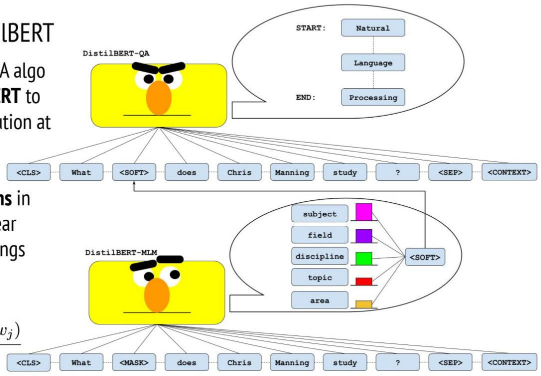
Our method: SCA with DistilBERT

2 modifications to the original SCA algo

- We use a **pre-trained DistilBERT** to output the probability distribution at a sampled token position
- We truncate the probability distribution to the **top k tokens** in the vocabulary and take a linear combination of their embeddings

Resulting soft token embedding:

$$E_s(x_t) = \frac{\sum_{j=0}^{k} p(w_j|x_{<t}, x_{>t})E(w_j)}{\sum_{j=0}^{k} p(w_j|x_{<t}, x_{>t})}$$



Related work

Other data augmentation methods for OOD QA [2]

- **Domain sampling**: determine which datasets can contribute more to OOD performance
- **Negative sampling**: include "no answer" segments and abstention option for model
- **Back-translation**: translate data to pivot language and back to target language
- **Active learning**: sample examples based on difficulty calculated by scoring functions

Data and setup

- **In-domain datasets**: SQuAD (Wikipedia), Natural Questions (Google queries on Wikipedia), NewsQA (news articles)
- **Out-of-domain datsets**: RACE (reading exams), DuoRC (movie reviews), RelationExtraction (synthetic relation questions)
- **QA task**: input: context paragraph and question; output: answer span
- **Models**: Pre-trained DistilBERT used as LM to generate soft tokens and as QA model to train on augmented data to perform QA task

Works Cited

1. Gao, F., Zhu, J., Wu, L., Xia, Y., Qin, T., Cheng, X., ... & Liu, T. Y. (2019). Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.*
2. Longpre, S., Lu, Y., Tu, Z., & DuBois, C. (2019). An exploration of data augmentation and sampling techniques for domain-agnostic question answering. *arXiv preprint arXiv:1912.02145.*

Experimental results

- Tuned hyperparameters to k = 5 and lr = 3e-5
- Across 3 masking strategies and varying % augmented, best model was masking 10% of tokens separately **F1: 49.2, EM: 32.2**, (baseline: F1: 47.72, EM: 30.63) (Fig 1)
- Improvement is specific to OOD dev sets (Fig 2), suggesting SCA improves robustness specifically
- Augmenting only context: similar improvements

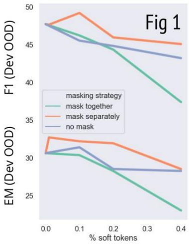


Fig 1

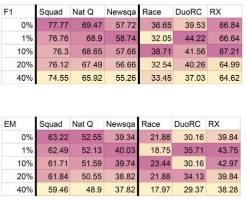


Fig 2

F1	Squad	Nat Q	Newsqa	Race	DuoRC	RX
0%	77.77	69.47	57.72	36.65	39.53	66.84
1%	76.76	68.9	58.74	32.05	44.22	66.64
10%	76.3	68.65	57.66	38.71	41.56	67.21
20%	76.12	67.49	56.66	32.54	40.26	64.99
40%	74.55	65.92	55.26	33.45	37.03	64.62

EM	Squad	Nat Q	Newsqa	Race	DuoRC	RX
0%	63.22	52.55	39.34	21.88	30.16	39.84
1%	62.49	52.13	40.03	18.75	35.71	43.75
10%	61.71	51.59	39.74	23.44	30.16	42.97
20%	61.84	50.55	38.82	21.88	34.13	39.84
40%	59.46	48.9	37.82	17.97	29.37	38.28

ID datasets OOD datasets

Analysis

Lower rate of complete misses

- Out of 382 dev examples, our model predicted the exact answer and baseline didn't for **25**, and **9** vice versa.
- Counted "complete misses" (CM): incorrect answers which had no **containment relation** with the correct answer.
- Out of 9 our model got wrong, **22%** were CM. Out of 25 baseline got wrong, **44%** were CM. Example of non-CM:

```
Context: NKG2D is encoded by KLRK1 gene which is located in the NK-gene complex (NKC)
situated on chromosome 6 in mice and chromosome 12 in humans.
Question: What is the name of the chromosome where you can find NKG2D?
Correct answer: Chromosome 12  Our model's answer: Chromosome 6 in mice and chromosome 12
```

Higher success rate on context-question pairs where paraphrasing is important

- Counted context-question pairs which had **paraphrasing** between context and question
- Out of 9 our model got wrong, **22%** had paraphrasing. Out of 25 baseline got wrong, **40%** had paraphrasing.
- To analyze the effectiveness of our soft tokens, we fed the **context** into the pre-trained DistilBERT we used in SCA to see if paraphrased words in the **question** were included in the truncated soft token distribution.

```
Context: Griffin gives them the ArcNet and explains it can only work in zero gravity: K
gets the idea to head to Cape Canaveral on 16th July 1969 (the day the Apollo 11 ship
launched).
Question: Where must they go to attach the ArcNet?
Correct answer and our model's answer: Cape Canaveral  Baseline's answer: No k
```

- In the example above, given the context with the token "`head`" masked, DistilBERT predicts: **`return (22.5%), fly (14.4%), travel (14.4%), go (9%), sail (5.7%)`**

Conclusion

Soft contextual data augmentation (SCA) **improves QA performance specifically on out-of-domain datasets**. Qualitatively, our results suggest that **SCA tends to improve performance when paraphrasing is involved** between the context and the question.