Robust QA System with Data Augmentation

Kai Ding
University of California, Berkeley
kai.ding@berkeley.edu

Abstract

In real-world situations, Question and Answering systems will rarely encounter data that look the same as training datasets. In this project, we explored the use of data augmentation techniques to improve the robustness of Q&A systems to domain shifts. In particular, we experimented with synonym replacement and back translation. Both techniques clearly improved baseline performance of EM and F1 scores, with synonym replacement achieving best results of **35.60 EM** and **51.03 F1 score** compared with baseline of 32.98 EM and 47.66 F1, although it is also sensitive to hyperparameter selection. We also discussed current challenges of applying data augmentation techniques to QA tasks, and invited further studies to advance the field.

1. Introduction

In recent years, QA systems have increasing applications in industries and daily settings, with examples such as digital assistants and customer service chatbots. But QA systems are so far highly dependent on training datasets, and generalization to unseen, or out-of-domain datasets provide an important gauge of whether such models truly understand the meaning of texts, and hence requires the system to model complex interactions between questions and context paragraphs

In existing research, data augmentation techniques have achieved some success with NLP tasks. EDA (easy data augmentation) was first proposed by Wei and Zou[1], and originally intended to apply to text classification tasks. Out of the 4 EDA techniques, we selected synonym replacement for experimentation, because we hypothesized that it is likely to be the most generalizable technique to QA task. The second data augmentation technique we experimented with was back-translation[2], and we further leveraged MarianMT[3], a NMT framework based on C++, for efficient computation.

The goal of this study was to experiment with the effectiveness of such techniques, shed light on practical challenges, and generate insights on how to advance the field to improve QA system robustness.

2. Project Overview

2.1 Datasets

We used 3 in-domain training datasets and evaluated performance on 3 out-of-domain training datasets. We also used a small number of examples (~127 each) from out-of-domain datasets for additional finetuning.

The 3 in-domain datasets include SQuAD[5] (Wikipedia paragraphs with crowd-sourced questions and answers) , NewsQA[6] (crowd-sourced based on CNN news), and Natural Questions[7] (Google search queries). The 3 out-of-domain datasets include DuoRC[8] (miscellaneous sources), RACE[9] (English language exams for middle and high school students) and RelationExtraction[10] (miscellaneous).

2.2 Approach

2.2.a Baseline model

We used DistilBert as the baseline model primarily due to computational considerations. According to the original paper[4], DistillBert is a faster and lighter version of Bert that has 40% less parameters, runs 60% faster while preserving over 95% of the BERT performance.

2.2.b Evaluation metrics

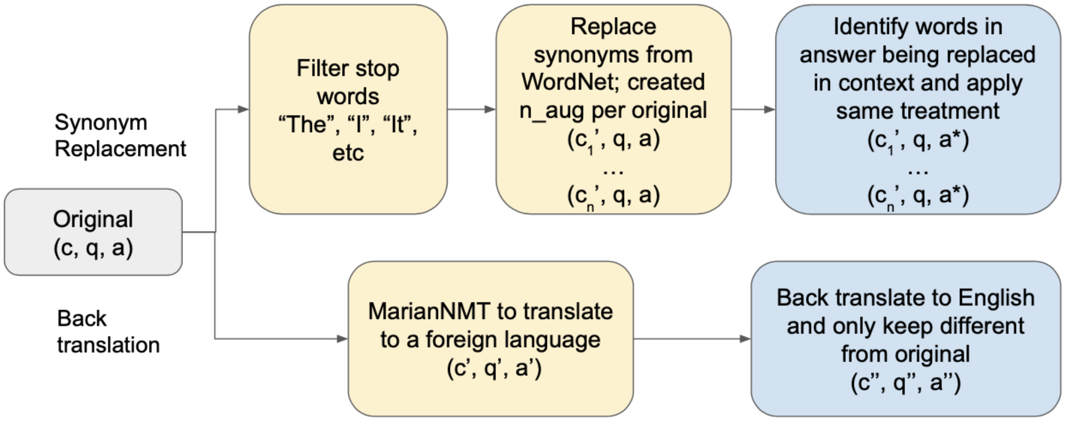
We used EM (Exact Match) and F1 score as evaluation metrics.

2.2.c Data processing and augmentation flow

Due to the difference in synonym replacement and back translation in how they alter texts, we applied different data processing techniques. For synonym replacement, after filtering for common stop words, we applied to context paragraphs. To keep answers in augmented context, we identified words that are being replaced in contexts, and applied the same replacement in answers. It's challenging to apply the same technique for back translation, due to words being potentially swapped, deleted as well as new filler words being introduced, and therefore we applied back translation on contexts as well as questions and answers.

Overall, with the data processing flows, we are able to preserve over 90% examples in context span for synonym replacement, while slightly over 50% examples for back translation.

Figure 1: Data processing & augmentation flow


Original
(c, q, a)
Synonym Replacement
Back translation
Filter stop words "The", "I", "It", etc
Replace synonyms from WordNet; created n_aug per original
(c_1', q, a)
…
(c_n', q, a)
Identify words in answer being replaced in context and apply same treatment
(c_1', q, a^*)
…
(c_n', q, a^*)
MarianNMT to translate to a foreign language
(c', q', a')
Back translate to English and only keep different from original
(c'', q'', a'')

2.2.d Translation frameworks

We experimented with two translation frameworks, including Python Back Translation 3.0 library which leverages Google Translate API and MarianMT which was proposed by Junczys-Dowmunt etc (2018)[4] as an efficient neural machine translation model based on C++. We discovered that Back Translation 3.0, though easy-to-implement, suffers from an hourly request limit due to Google API constraints. Therefore, we used MarianMT to perform all back translations for all of our experiments.

2.2.e Hyperparameter tuning

Due to computational considerations, we only performed hyperparameter tuning on learning rate, adjusting it from 3e-5 default value to 1e-5, because we saw clear signs of overfitting when training with a higher learning rate on the much smaller out-of-domain training sets.

3. Experiments

3.1 Results

Overall, synonym replacement with hyperparameters p (% of words being replaced) = 0.3, and n_aug (number of augmented examples per original) = 9 achieved the best performance of **35.60 EM** and **51.03 F1 score**. Further breakdown of the experiments are included in the following sections.

However, we'd like to point out at least two limitations in our experiments. Firstly, due to large computational costs of back translation, where it took on average 3-5 hours on GPU to process just ~500 examples in one language, the back translation results are achieved on much fewer augmented examples compared with synonym replacement, which on default value n = 9 generates 8 new augmented examples, whereas back translation only generates 1 new augmented example per original. Secondly, also due to computational and time limits, we did not perform extensive novel combinations of techniques.

Figure 2: Experiment results summary

	Model	**Out-of-domain validation scores**	
		EM	**F1**
Baseline	#1: DistillBert trained on in-domain data	32.98	47.66
	#2: DistillBert further trained on out-of-domain training data	31.94	48.50
Synonym Replacement	p = 0.05, n = 9	34.03	49.36
	p = 0.05, n = 18	34.03	49.65
	p = 0.1, n = 9	34.29	48.81
	p = 0.1, n = 18	33.76	49.81
	p = 0.2, n = 9	33.76	49.77
	p = 0.2, n = 18	35.34	50.76
	p = 0.3, n = 9	35.60	51.03
	p = 0.3, n = 18	34.55	50.24
Back Translation	En -> Es -> En	34.03	50.00
	En -> Zh -> En	34.29	49.47
SR + BT	SR (p=0.3,n=9) -> Es -> En	33.76	49.93

3.2 Baseline model

We created two baselines with DistilBert, with baseline 1.0 fine tuned on in-domain datasets and baseline 2.0 further fine tuned on the small samples of out-of-domain datasets, and both evaluated against the in-domain and out-of-domain validation sets. Not surprisingly, DistillBert, while achieving **54.25 EM and 70.39 F1 score** on in-domain validation sets, suffered significant drop to **32.98 EM and 47.66 F1 score** when evaluated against the out-of-domain validation sets, meaning the model does not generalize well to data it has not seen before. When further fine tuning on the out-of-domain training set, due to the small sample size, there was no clear improvement.

3.3 Synonym replacement

We experimented with a range of hyperparameter values, including p (% of words being replaced) and n (number of augmented examples per original), where p = {0.05, 0.1, 0.2, 0.3} and n = {9, 18} for a total of 8 experiments. Whereas in the original EDA paper p = 0.1 achieved the best results for text classification tasks, our experiments showed p = 0.3 achieved better performance for QA tasks, potentially indicating QA tasks require more data augmentation than text classification. In particular, p = 0.3 and n = 9 achieved the best performance in our experiments; when n = 18, the model started to regress in performance, potentially indicating overfitting.

3.4 Back translation

We experimented with two very different languages as the middle language of back translation, including Spanish and Chinese, under the hypothesis that languages with different syntax and structure might produce different results. But in our limited experiments, both achieved very similar results.

3.4 Mixed techniques

We combined the best-performing synonym replacement parameters (p=0.3, n=9) with back translation (en -> es -> en), but there appeared to be no sign of improving performance with our limited experiments.

4. Case Analysis

4.1 Robustness through synonym replacement

Out of 382 validation examples, the best-performing model by synonym replacement (p=0.3, n=9) differed from baseline in 75 examples. After reviewing all the disagreed examples, we hypothesized that SR introduced robustness potentially through two ways, firstly by helping the QA system handle difficult vocabulary, and secondly, perhaps surprisingly, by helping the QA system navigate sentence structure better. We demonstrated the common themes with 2 examples below, which are hardly isolated.

Example 1: “schutterstuk” means group portrait. We hypothesized that, given words with similar meaning including “paint”, “picture” and “portrait” appeared in training set, synonym replacement potentially helped QA system approximate meaning of “schutterstuk” or ““schutterstuk painted” better than baseline, and therefore correctly distinguishes it from a location.

- **Context:** The Banquet of the Officers of the St George Militia Company in 1627 refers to a schutterstuk painted by Frans Hals for the St. George (or St. Joris) civic guard of Haarlem, and today is considered one of the main attractions of the Frans Hals Museum there.
- **Question**: What is the name of the place where The Banquet of the Officers of the St George Militia Company in 1627 can be found?
- **Baseline answer:** schutterstuk
- **Best model answer:** Frans Hals Museum

Example 2: Baseline model was tricked by the semantic relationship introduced by “but”, while best model correctly identifies the answer after “but”. This is surprising because intuitively synonym replacement dealt with word meanings rather than semantic structure.

- **Context:** The Consecration of Aloysius Gonzaga as patron saint of youth is a c.1763 painting attributed to Francisco de Goya and now owned by the town of Jaraba but stored in the Saragossa Museum in Saragossa.
- **Question**: What is the name of the place where Consecration of Aloysius Gonzaga as patron saint of youth can be found?
- **Baseline answer:** Jaraba
- **Best model answer:** Saragossa Museum

4.2 Nonsensical synonym replacements

When increasing hyper parameters of synonym replacement to generate more augmented examples, there was a tendency to produce more nonsensical synonyms, resulting in the model to overfit.

Example: Synonym replacement mistakenly changed the meaning of “operation” and produced replacements that made little sense and took the word out of context.

- **Original:** Riseberga Abbey (Swedish: Riseberga kloster), was a Cistercian nunnery in Sweden, in operation from cirka 1180 until 1534
- **Augmented #1:** riseberga abbey swedish riseberga kloster was a trappist nunnery in kingdom of sweden in mental process from cirka 1180 until 1534
- **Augmented #2:** riseberga abbey swedish riseberga kloster was a trappist nunnery in kingdom of sweden in surgery from cirka 1180 until 1534
- **Augmented #3:** riseberga abbey swedish riseberga kloster was a trappist nunnery in kingdom of sweden in military operation from cirka 1180 until 1534
- **Augmented #4:** riseberga abbey swedish riseberga kloster was a trappist nunnery in sverige in mathematical operation from cirka 1180 until 1534
- **Augmented #5:** riseberga abbey swedish riseberga kloster was a trappist nunnery in sverige in surgical procedure from cirka 1180 until 1534

5. Conclusion

This project demonstrated the effectiveness of data augmentation techniques for improving QA robustness, where both synonym replacement and back translation visibly improved baseline performance. However, we have identified several topics for further research to apply data augmentation at scale.

Firstly, we can consider techniques to identify answer span in context, even without the exact same wording; this matters especially for techniques such as back translation. The technique needs to recognize words that matter most to the meaning of the answer, and identify the most similar words in context. The technique also needs to handle changing word orders and new filler words introduced via translation.

Secondly, EDA (easy data augmentation) techniques are highly dependent on specific NLP task and training datasets which might have different lengths of contexts. It is challenging to apply EDA techniques more broadly, because techniques such as synonym replacement does not consider context and is liable to alter the original meaning of the sentence. We believe back translation is potentially more generalizable, provided we solve for the first challenge and have efficient ways of computation.

Finally, EM and F1 scores are highly imperfect measures for QA tasks, given in human language even a question can be ambiguous and has no single correct answer. We expect new metrics to be developed that shed more insight on what it really means to understand.

References

[1] Jason W. Wei and Kai Zou. EDA: easy data augmentation techniques for boosting performance on text classification tasks. CoRR, abs/1901.11196, 2019.

[2] Yu et al. QANet: Combining local convolution with global self-attention for reading compre- hension. In arXiv preprint arXiv: 1804.09541, 2018.

[3] Junczys-Dowmunt et al. Marian: Fast Neural Machine Translation in C++. In arXiv represent arXiv: 1804.00344, 2018

[4] Sanh et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In arXiv preprint arXiv:1910.01108, 2019

[5] Pranav Rajpurkar et al. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In arXiv preprint arXiv: 1606.05250v3, 2016

[6] Adam Trischler et al. NewsQA: A Machine Comprehension Dataset. In arXiv preprint arXiv: 1611.09830v3, 2017

[7] Amrita Saha et al. DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension. In arXiv preprint arXiv: 1804.07927v4, 2018

[8] Guokun Lai et al. RACE: Large-scale ReAding comprehension dataset from examinations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 785-794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[9] Omer Levy et al. Zero-Shot Relation Extraction via Reading Comprehension. In arXiv preprint arXiv: 1706.04115v1, 2017