

DA-Bert: Achieving Question-Answering Robustness via Data Augmentation

Stanford CS224N Default Project

Andre Turati, Jose Francisco, Vineet Edupuganti

Department of Computer Science

Stanford University

{amturati,josencf,ve5}@stanford.edu

Abstract

Pretrained models are the basis for modern NLP Question-Answering tasks; however, even state-of-the-art models are heavily influenced by the datasets they were trained on and don't generalize well to out-of-domain data. One avenue for improvement is augmenting the training dataset to include new patterns that may help the model generalize outside of its original dataset. In this paper, we explore improving model robustness in the question-answering task via various data augmentation techniques including adding noise to our contexts and backtranslating (translating text to a pivot language and then back) both the queries and contexts. We find that utilizing the technique of backtranslation on the queries, both on in-domain and out-of-domain training datasets, greatly improves model robustness and gives 3.7% increase in F1 scores over our baseline model from 48.43 F1 to 50.24 F1 in the validation set and a 59.27 F1 score on the test set. Further, within this approach of backtranslation, we explore the linguistic effect of particular pivot languages and find that using Spanish adds the greatest robustness to our model. We theorize that Spanish and potentially other Romance languages' linguistic similarity to English gives clearer and more helpful translations than other high-resource languages with different roots.

Mentor: Davide Giovanardi

1 Introduction

One of the biggest problems in NLP Question-Answering systems is their inability to scale and adapt well outside of the data they were trained in. Concretely, they perform well on *in-domain* datasets but perform poorly on *out-of-domain* datasets. We propose a system that pushes the boundary of performance of out-of-domain question answering. This system can be applied to question-answering tasks that generalize to different domains such as semantic search or medical analysis (e.g. a doctor searching "what drug is best for X").

To accomplish this, we propose DA-Bert: an extension of a DistilBert Model improved via data augmentation to perform the task of robust and adaptable question-answering to out-of-domain data.

The question-answering task is framed as such: given a context passage c , a query about the context q , and a span of text from the context corresponding to the answer a , our model aims to select a span of text that matches a as closely as possible. Concretely, our model will select a start and end location within the passage corresponding to what it predicts for the answer span. To measure similarity to the gold answer and train our model, we aim to minimize the the sum of the negative log-likelihood (cross-entropy) loss for the start and end locations, i and j respectively, of the gold answer:

$$L(\theta) = -\log p_{start}(i) - \log p_{start}(j)$$

We finetune our models on in-domain data (examples taken from the SquAD, NewsQA, and Natural Questions datasets) [1, 2, 3] and evaluate our models according to F1 and EM metrics between the true answer a and our model's selected answer span for out-of-domain data (examples taken from the DuoRC, RACE, and RelationExtraction datasets) [4, 5, 6].

We start with a pretrained DistilBert model as our baseline and aim to improve over this model [7]. To do so, we utilize data augmentation techniques in order to produce novel training samples that may help the model generalize to out-of-domain data. We focus on two main augmentation techniques: adding sampled noise from the context and backtranslation of both context and query to produce augmented examples.

We find that backtranslation techniques work best among our approaches; in particular, backtranslation of the query to get a new modified question shows the largest performance gains in F1 scores (over 3.7% F1 improvement over the baseline model). We further explore the significance of the intermediate pivot languages in back-translation and find that using Spanish gives the largest gains in F1 scores; we hypothesize that Spanish and potentially other Romance languages’ grammatical and linguistic similarity to English (as suggested by [8]) leads the backtranslation to produce translations with higher fidelity to the original text.

2 Related Work

2.1 Adversarial Training

There has been much research into data augmentation techniques for NLP systems. One area of research that is particularly relevant to model robustness is adversarial training presented by [9]. The researchers in [9] exploit model weaknesses by creating ‘adversarial’ examples with the intent of tricking the QA model. Particularly relevant to our approach is their use of concatenative data augmentation techniques: starting with a query, context, answer triple (q, c, a) , they append different types of noise s to the context in order to obtain the augmented example $(q, c + s, a)$. Various concatenations were tried in the paper ranging from appending an entire sentence to sampling random tokens from the vocabulary. They show that these small modifications drastically degrade performance on F1 scores. We build upon these adversarial training methods to explore how smaller changes can actually increase performance of the model on out-of-domain data.

2.2 Backtranslation Augmentation

Extensive work has been done in exploring backtranslation as a data augmentation technique, particularly in neural machine translation for low-resource languages. [10] shows that using back-translation to augment the corpus to create more data for Neural Machine Translation (NMT) significantly improves BLEU score performance on translation tasks. Moreover, [11] explores the use of different pivot languages and replacement of parts of the corpus with their translations in other languages. They find that high-resource languages (such as Spanish, German, French, etc.) outperform low-resource languages like Swahili or Urdu as augmentors. We utilize these backtranslation techniques that generally are applied to NMT in our question-answering task to create paraphrased examples. Further, we explore if there is a correlation between pivot languages and model performance in our question-answering task when performing backtranslation with a select number of languages. [8] suggests that languages that are grammatically and linguistically similar to English (e.g. Spanish) could potentially produce higher quality backtranslations with high fidelity in intent and meaning to the original text.

3 Approach

In this section, we describe the four data augmentation strategies we introduce to evaluate model robustness of the baseline DistilBERT model (with our control being no data augmentation). That is, our baseline EM and F1 scores are that of the default DistilBERT model with the original data [7].

Starting with our original training dataset $\mathcal{D}$, we randomly sample a subset d that we will augment. As Figure 1 shows, starting with a training example $x_i \in d$, we produce a new augmented example $x'_i = f(x_i)$ utilizing the augmentation functions f described in the following subsections. Finally, we get $d' = \{f(x_i) \mid x_i \in d\}$ and concatenate this with $\mathcal{D}$ to get our new training dataset $\mathcal{D}' = \{\mathcal{D}, d'\}$.

We experimented with various sizes of d across the different strategies and found that in all cases $|d| = .2 * |\mathcal{D}|$ (i.e. augmenting 20% of the original training examples) resulted in the best out-of-domain performance.

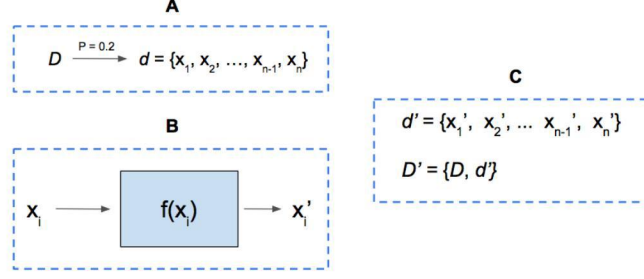


Figure 1: Methodology for augmenting training data to enhance robustness, shown in three phases. A) A set of examples d is sampled from the original training data D with probability 0.2. B) Augmented training examples $x'_i = f(x_i)$ are generated from sampled training examples $x_i \in d$. C) The final dataset D' is produced by merging the original and augmented data.

3.1 Add Random Noise

The first data augmentation technique we explore is adding random noise to the end of the context for each training example. The inspiration for this method is prior work focused on adversarial manipulation of training data [9]. Because the objective in this case is augmentation with the intent of improving performance, we opt for a simple and minor augmentation strategy that is less dramatic than the methods presented in literature on adversarial modifications.

Specifically, our differentiated approach involves a series of tokens being appended to the end of the context string for various examples during data processing to create new examples. Additionally, each appended portion is obtained by selecting random words from the original context itself.

Mathematically, starting with a query, context, answer triple $x_i = (q, c, a)$ we randomly sample a set of tokens $s = \{t_1, \dots, t_n\}$ from the original context c ; we then use this to obtain our augmented example $x'_i = f_{add}(x_i) = (q, c + s, a)$.

The motivation behind this approach is that appending to the end of the context is less likely to distort meaning (as oppose to breaking semantic patterns in the middle of the context), and the use of words from the context itself provides a greater challenge than inserting random tokens from the entire vocabulary the model may learn to ignore.

For each example in the training data, we generate these altered counterparts with appended text to the context passage. We define the length of the appended portion to be a random value l between 5% and 10% of the original context length. This range was validated with some initial experiments, but we leave more thorough manipulation of this hyperparameter for future work.

3.2 Context Backtranslation

Additionally, we explore backtranslation (translating from English to an intermediate pivot language and back again) by leveraging the BackTranslation Python library [12]. With this approach, we “paraphrase” each training example to produce semantically equivalent content that is textually different for the model to learn from. This is a concept that has been explored for different applications in prior work [10, 13, 14].

In our case, we first explore backtranslation of the context only (for a fixed query). This method has been less widely studied in the literature per our review. Concretely, starting with a query, context, answer triple $x_i = (q, c, a)$ and a pivot language L , we obtain $c' = \text{backtranslate}(c, L)$ where backtranslate is a function that translates c from English to L and then back to English. Then, we obtain our augmented example $x'_i = f_{trans-context}(x_i) = (q, c', a)$.

To test the context backtranslation approach, we implement an algorithm that translates a given context into various languages using [12] and then back to English for 20% of randomly selected training examples in the original dataset D . For the newly generated examples, we check if the original answer still appears after the backtranslation process. If so, we then search for the index of the answer in the backtranslated context paragraph and set the new answer start at the index of the first occurrence of the

answer. Note that this procedure only adds backtranslated examples to our dataset when the answer remains in its original form - without this simplification, we would have to perform semantic matching to find the correct answer span in the new text (since the exact verbiage would be different), which could be prone to errors. We observe some minor differences in the quantity of augmented data with this method based on the pivot language (see description of future work in the Conclusion for ideas for next steps).

We perform experiments on five different languages: Spanish, German, Chinese, Russian, and Arabic. Spanish and German were chosen due to their linguistic similarity to English (especially for the former) according to [8]. The remaining languages were chosen due to their lexical differences (dissimilar alphabets, pronoun usage, morpheme combination patterns, etc.) [15]. All results and further analysis of the difference in performance for these languages are described in the Results section.

3.3 Query Backtranslation

In a similar vein, we also explore backtranslation of the query with a fixed context to create augmented examples. Concretely, starting again with our example $x_i = (q, c, a)$ and a pivot language L , we obtain $q' = \text{backtranslate}(q, L)$. Then, we obtain our augmented example $x'_i = f_{\text{trans-query}}(x_i) = (q', c, a)$.

With this setup, we keep all augmented examples (corresponding to 20% of randomly selected training examples as described earlier) because there is no risk of losing the answer span as in the case for context translation. We use the same set of languages for this analysis: Spanish, German, Chinese, Russian, and Arabic.

3.4 Out of domain finetuning and augmentation

Finally, to boost performance on our out of domain validation and test sets, we perform a finetuning step on the out-of-domain train set for the best performing model configurations. To further increase performance, we augment this limited out of domain data via the query backtranslation approach (with Spanish as the pivot language) to create 20% more samples. This practical modification provided us with additional gains on the out-of-domain validation set.

4 Experiments

This section describes the data we use to train and test our model. Further, we describe the quantities we used to evaluate the performance of our model. Finally, we go into detail on the experiments ran using the add noise augmentation method, context backtranslation, query backtranslation, and various hyperparameter changes such as changing the learning and dropout rates of the default DistilBERT model.

4.1 Data

We experiment using the in-domain training set and the in-domain validation set. Specifically, for our question-answering task, we use the SQuAD [1] and Natural Questions [3] datasets originating from Wikipedia and NewsQA [2] dataset as our in-domain datasets (150K train examples and 27.5K validation examples collectively). For out-of-domain datasets used to evaluate model robustness, we use DuoRC [4] dataset from movie reviews, RACE [5] from an examinations dataset, and RelationExtraction [6] from Wikipedia articles (381 train examples and 382 validation examples collectively). For our best configurations, we also finetune on the augmented out-of-domain train set described in Section 4.4 to achieve a performance boost.

4.2 Evaluation method

To evaluate our experimental results, we compare the EM and F1 scores of our experimental models with those of the baseline. Specifically, the baseline DistilBERT model (which uses 6 layers, 12 attention heads, a dropout rate of 0.1 and an attention dropout rate of 0.1) produces an F1 score of 70.95 and an EM score of 55.07 on the in-domain validation dataset. Further, it produces an F1 of

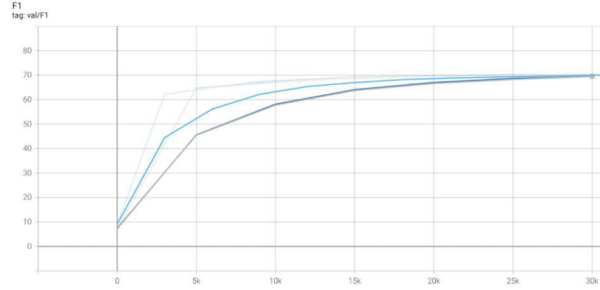


Figure 2: Pictured in light blue is the *in-domain* validation F1 score over time for the best version of query backtranslation (using Spanish as the pivot language). The gray line represents the F1 score for adding noise, which overlaps with the darker blue line for context backtranslation (and therefore is not easily visible in plot). The faint, transparent curves in the plot are Tensorboard artifacts associated with the smoothing that was used for clearer visualization.

48.43 and an EM of 33.25 in the out-of-domain validation dataset. We aim to achieve a higher F1 and EM score using the forms of data augmentation specified in the Approach section above.

4.3 Augmentation Experiments

We train models for each of the primary augmentation techniques described above: adding noise, context backtranslation, and query backtranslation. Of those three, we found query backtranslation to be the most promising of the experiments (see details in the Results section). We then seek to determine the effects of different pivot languages on the query back-translation. That is, we train query backtranslation models using each of the following languages: Spanish, Arabic, Russian, Chinese, and German and compare results across them. Furthermore, using Spanish as the pivot language for query backtranslation (our best result), we change the configuration and the learning rates of the default DistilBERT model to optimize performance. We ran two trials for these experiments to assess variability and the impact of the selected examples for backtranslation, which yielded no significant difference in results. More detailed analysis of variability was limited by lack of computational resources.

4.4 Model Configuration Experiments

Additionally, we carry out an experiment on 10 percent of the data to search for the best model configurations, training eight different models with modified configurations (four parameter combinations for each of the baseline and the query backtranslation with Spanish). Specifically, we adjusted the number of layers, number of heads, and the dropout rates of the DistilBERT model. In these experiments, which are shown in Table 2, the dropout rate and the attention dropout rates were kept equal. Note that an assumption behind this approach is that optimal model configurations (as measured by out-of-domain validation performance) on 10 percent of the data extend to the full dataset. This is a simplification that was made to save compute time, but in an ideal scenario this grid search would have taken place on models trained on the full data.

5 Results

We discuss four different results. First, we compare the F1 scores of our three data augmentation methods and reasons for their success/failures. We also compare the best of those three robust models (the query backtranslation) to the baseline. Finally, we observe the impacts of both varying the query-backtranslation pivot language and other DistilBERT configuration parameters.

5.1 Augmentation Techniques

Of the three augmentation techniques, backtranslating the query (with Spanish as the pivot) outperformed the rest, with an F1 score of 48.74 on the OOD validation set (before optimizing further or finetuning on the augmented OOD train set, which both contributed substantially to end performance). Meanwhile, translating the context (with Spanish as the pivot) and adding noise with the method

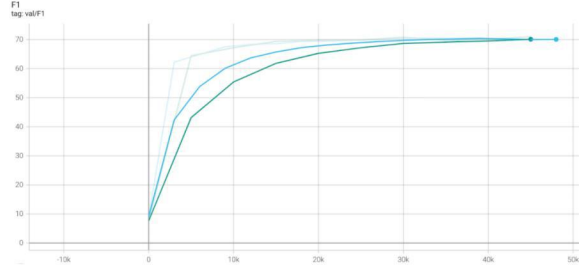


Figure 3: Pictured in light-blue is the *in-domain* F1 score over time for the strongest version of query backtranslation (using Spanish as the pivot language). The green line represents the baseline DistilBERT model’s F1. Note the faster rise in the F1 score for our augmented query backtranslation approach (with Spanish) throughout training. The faint, transparent curves in the plot are Tensorboard artifacts associated with the smoothing that was used for clearer visualization.

described previously produced F1 scores of 48.33 and 47.88 respectively. We can see that query backtranslation learned more effectively than the other two strategies which were similar in performance on the in-domain validation F1 curves (Figure 2).

Query backtranslation likely did the best for a few reasons. First, queries are generally short (< 10 words total). Thus, there is less room for meaning to get distorted as compared to longer contexts. Additionally, the shorter paraphrases were enough to give the model a different signal with the same intent as the original query.

There are various hypotheses as to why the other two strategies were not as effective as query backtranslation. For our added noise augmentation, we hypothesize that random tokens drawn from the context may in fact be too adversarial for the model to ignore, and this may disrupt the model’s training process. For the context backtranslation approach, we hypothesize that firstly, contexts are much longer than the query, so there is a larger probability that the meaning of parts or all of the context is heavily distorted during the translation process. Secondly, our method of reselecting the answer span in the backtranslated context is prone to errors since we are selecting the first occurrence of the original answer if it is present, whether or not it matches in position in the new context to its position in the original context.

5.2 Query Backtranslation vs Baseline

Comparing the query backtranslation with the baseline, we see large improvements in the in-domain validation performance. As seen in Figure 3, the query backtranslation using Spanish as a pivot reached higher F1 scores than the baseline model at each time-step until they both plateaued at around 30k iterations. The final F1 score on the OOD validation set of the query backtranslation also outperforms the baseline on the out-of-domain validation set (F1 baseline score of 48.432 compared to query backtranslation score of 48.74). To boost this score further, we tuned the model architecture as shown in Section 6.4 and also augmented the out of domain train set by 20% as described in Section 4.4 (finetuning for 5 epochs). These approaches yielded our best F1 result of 50.242.

5.3 Pivot Languages for Query Backtranslation

Further, we wished to explore the effects of different pivot languages during backtranslation. Testing five different backtranslation languages, we see in Table 1 that changing the pivot language also changes the final F1 score of the model. That is, backtranslation pivot language is a non-trivial hyperparameter. Out of the five languages tested, we find that Spanish outperforms the other languages in EM and F1 across in-domain and out-of-domain datasets with the exception of German on the in-domain F1 scores. As mentioned previously, we hypothesize that German and especially Spanish’s linguistic similarity and shared roots produce more coherent and usable backtranslations than the other languages tested.

5.4 Query Backtranslation Model Configuration and Hyperparameters

Finally, in Table 2 we examine the effects of two chosen hyperparameters: dropout rate and number of layers of the DistilBERT model. We see that dropout rate had negligible effect on EM and F1 scores perhaps due to the deeper number of layers. Overall, the greatest improvement to the baseline

F1 + EM scores for various languages on query backtranslation				
Language	In-domain F1	In-domain EM	Out-of-domain F1	Out-of-domain EM
Spanish	70.32	54.49	48.74	33.25
German	70.60	54.41	48.37	33.25
Russian	69.05	53.25	45.24	29.84
Arabic	68.98	52.91	45.99	30.63
Chinese	69.76	53.73	45.78	29.84

Table 1: EM and F1 scores for different backtranslation languages on a backtranslate-query model using the default DistilBERT configuration. We chose a Romance language (Spanish), a Germanic language (German), a Slavic language (Russian), a Semitic language (Arabic), and a Sino-Tibetic language (Chinese). Bolded are the highest scores for each dataset.

F1, EM scores for 10% of training data				
Configuration Setup	In-domain Baseline (F1,EM)	OOD Baseline (F1,EM)	In-domain Translate Query (F1,EM)	OOD Translate Query (F1,EM)
6 Layers, Dropout 0.1 *	56.71, 38.59	37.08, 20.39	61.87, 45.28	39.77, 23.30
8 Layers, Dropout 0.05	61.41, 45.61	37.69, 21.20	61.93, 46.85	37.27, 21.99
8 Layers, Dropout 0.2	61.41, 45.61	37.69, 21.20	61.93, 46.85	37.27, 21.99
12 Layers, Dropout 0.05	61.12, 45.05	38.06, 23.56	63.04, 48.02	48.51, 34.29
12 Layers, Dropout 0.2	61.12, 45.05	38.06, 23.56	63.04, 48.02	48.51, 34.29

Table 2: F1, EM scores for various configurations, each trained on 10% of the in-domain data. We raise the number of layers from the default of 6. Note that the number of attention heads is always double the number of layers. Dropout rates apply to both general dropout and attention dropout. Bolded are the highest scores for each dataset. The backtranslation language used is Spanish.

* *Default*

comes from increasing the number of layers to 12. Note that in this experiment, we use Spanish as our backtranslation language and we use an optimal learning rate of $1e - 5$.

6 Analysis

As mentioned above in the Results, the Spanish backtranslation model outperforms the other augmentation techniques in F1 on our OOD validation set. However, we still wish to examine where this model fails and on what kinds of datapoints we can improve on in future work.

To perform a qualitative analysis of our model, we load in our training checkpoints for query backtranslation with Spanish (our best augmentation configuration as described earlier) and use Tensorboard to look closely at model outputs for a fixed holdout set (comprised of in domain train/val and out of domain train/val examples).

Across model configurations and examples, we discover that the model struggles to effectively handle long sequences. This is a result of the fact that there is a large quantity of information embedded within these contexts, making it hard to determine dependencies between words and phrases. The example below shows all mentions of time in the context for the given "when?" question. In this case, the predicted answer was completely unrelated, although it appeared near "1991" (also incorrect, but denoting a time).

Query: When was her surgery?

Context: Natalie Cole’s search for a new kidney ended this week when someone with a compatible organ died...Cole, who underwent a successful kidney transplant at Cedars-Sinai Medical Center in Los Angeles **Tuesday**, was "responding well and recuperating" Thursday, her publicist said. ...Cole’s need for a kidney donation was highly publicized, including with an appearance on

CNN's "Larry King Live" on March 31... for her 1991 critically acclaimed album "Unforgettable: With Love," a jazzy tribute to her father...Cole said her kidney troubles date to February 2008...

Actual Answer: Tuesday

Predicted Answer: acclaimed album "Unforgettable: With Love," a jazz

In general, however, the "when" questions had reasonably effective performance with query backtranslation for shorter sequences, even when this process yielded an imperfect paraphrase. The result below is one example of this phenomenon and indicates that the model looks to pair the "when" query with date phrases even when the backtranslated query by itself does not make complete sense.

Query: When was Marlena owned on days of our lives?

Context: In 1995, Marlena was featured in a supernatural possession storyline. The storyline caused huge controversy and garnered a lot of attention to the soap . At the time , the head writer was James E. Reilly , known for outrageous storylines .

Actual Answer: 1995

Predicted Answer: In 1995

Finally, another non-trivial trend we observed was the conversion of questions to statements after backtranslation, as in the following example, which is partially due to the backtranslation API we leveraged. Notably, we found no difference in the predicted answer for backtranslation as compared to the original query across our dataset for these circumstances. This similarity in result is likely because the sequence of words in the query encodes roughly similar semantic content regardless of whether it is posed as a question or a statement. In the example below, note also that the predicted answer is just as strong as the actual answer (in reality, both are incomplete).

Query: Types of social welfare programs in the United States.

Context: Social programs in the United States... include cash assistance , healthcare and medical provisions , food assistance , housing subsidies , energy and utilities subsidies , education and childcare assistance , and subsidies and assistance for other basic services . Private provisions from employers , either mandated by policy or voluntary , also provide similar social welfare benefits .

Actual Answer: housing subsidies

Predicted Answer: cash assistance , healthcare and medical provisions

7 Conclusion

Generalizing question answering systems to attain high performance on out-of-domain data that they have not been trained on is a substantial challenge in the field of NLP. With a set of comprehensive experiments, we demonstrate the merits of a data augmentation approach to mitigate these issues. Specifically, we find that performing English to Spanish query backtranslation with optimized hyperparameters (including number of examples to backtranslate) leads to an F1 score improvement of 1.81 over the baseline methods (3.7% rise) for the out of domain validation set and leaderboard. While this result is encouraging, a limitation of this work is that we did not experiment with backtranslating the query and context jointly to generate more diverse examples due to practical constraints. Such a method could improve the variety in the training set and improve generalization across out of domain datasets.

Furthermore, in the future, it would be valuable to perform context backtranslation without having to discard examples that lose the verbatim gold answer. Human labeling or rule-based heuristics (which would be cheaper) could be an effective way to carry out this process and produce the correct answer spans, which would then make this technique more powerful.

References

- [1] Konstantin Lopyrev Pranav Rajpurkar, Jian Zhang and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. *CoRR, abs/1606.05250*, 2016.
- [2] Xingdi Yuan Justin Harris Alessandro Sordoni Philip Bachman Adam Trischler, Tong Wang and Kaheer Suleman. Newsqa: A machine comprehension dataset. *ACL 2017, page 191*, 2017.
- [3] Olivia Redfield Michael Collins Ankur Parikh Chris Alberti Danielle Epstein Illia Polosukhin Matthew Kelcey Jacob Devlin Kenton Lee Kristina N. Toutanova Llion Jones Ming-Wei Chang Andrew Dai Jakob Uszkoreit Quoc Le Tom Kwiatkowski, Jennimaria Palomaki and Slav Petrov. Natural questions: a benchmark for question answering research. *Association for Computational Linguistics (ACL)*, 2019.
- [4] Mitesh M. Khapra Amrita Saha, Rahul Aralikkatte and Karthik Sankaranarayanan. Duorc: Towards complex language understanding with paraphrased reading comprehension. *ACL*, 2018.
- [5] Hanxiao Liu Yiming Yang Guokun Lai, Qizhe Xie and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. *EMNLP*, 2017.
- [6] Eunsol Choi Omer Levy, Minjoon Seo and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv:1706.04115*, 2017.
- [7] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- [8] Foreign Service Institute. Foreign language training. 2021.
- [9] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.
- [10] Michael Auli David Grangier Sergey Edunov, Myle Ott. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*, 2018.
- [11] Nitish Shirish Keskar Caiming Xiong Richard Socher Jasdeep Singh, Bryan McCann. Xlda: Cross-lingual data augmentation for natural language inference and question answering. *arXiv preprint arXiv:1905.11471*, 2019.
- [12] <https://pypi.org/project/BackTranslation/>.
- [13] Amane Sugiyama and Naoki Yoshinaga. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, 2019.
- [14] Wee Chung Gan and Hwee Tou Ng. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075, 2019.
- [15] Ying Wang and Jing Chen. Differences of english and chinese as written languages and strategies in english writing teaching. *Theory & Practice in Language Studies*, 3(4), 2013.

A Appendix

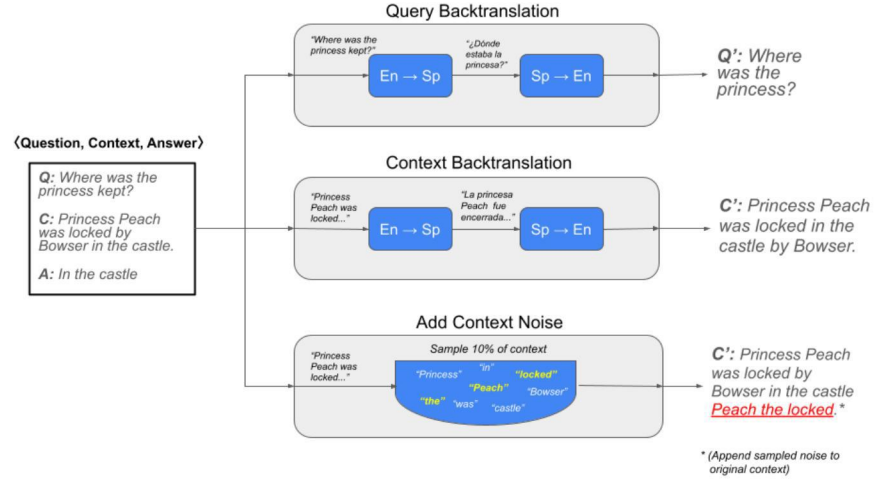


Figure 4: A visualization of our augmentation techniques on a simple example.