Data Augmentation for Biomedical Factoid Question Answering

Dimitris Pappas[1,2], Prodromos Malakasiotis[1,3], and Ion Androutsopoulos[1]

[1]Department of Informatics, Athens University of Economics and Business, Greece
[1]pappasd@aueb.gr,rulller@aueb.gr,ion@aueb.gr
[2]Institute for Language and Speech Processing, Research Center 'Athena', Greece
[2]dpappas@athenarc.gr
[3]Institute of Informatics and Telecommunications, NCSR 'Demokritos', Greece

Abstract

We study the effect of seven data augmentation (DA) methods in factoid question answering, focusing on the biomedical domain, where obtaining training instances is particularly difficult. We experiment with data from the BIOASQ challenge, which we augment with training instances obtained from an artificial biomedical machine reading comprehension dataset, or via back-translation, information retrieval, word substitution based on WORD2VEC embeddings or masked language modeling, question generation, or extending the given passage with additional context. We show that DA can lead to very significant performance gains, even when using large pre-trained Transformers, contributing to a broader discussion of if/when DA benefits large pre-trained models. One of the simplest DA methods, WORD2VEC-based word substitution, performed best and is recommended. We release our artificial training instances and code.

1 Introduction

Question Answering (QA) systems aim to answer natural language questions by searching in structured (Fu et al., 2020; Luo et al., 2018; Yadati et al., 2021) or unstructured data, such as free-text documents (Aghaebrahimian, 2018). Here we consider QA of the latter kind. Fully fledged QA systems for document collections retrieve relevant documents, identify relevant passages, extract and aggregate answer spans etc. (Chen et al., 2017a; Pappas and Androutsopoulos, 2021). There are also different types of questions, e.g., *yes/no*, *factoid*, *list*, *how-to*. Thus, creating realistic datasets to train and evaluate complete QA systems for document collections is resource intensive, especially for systems targeting specialized domains. A prime example is the *biomedical domain*, the focus of this work, where obtaining realistic training (and test) instances requires medical expertise, which is costly and difficult to obtain. Consequently, biomedical datasets for full QA systems contain just a few thousand training instances (Tsatsaronis et al., 2015; Möller et al., 2020) or focus on simpler question types only, e.g., yes/no questions (Jin et al., 2019).

A simplified form of QA for textual data is Machine Reading Comprehension (MRC) (Yang et al., 2015; Rajpurkar et al., 2016; Campos et al., 2016; Chen et al., 2017b; Lai et al., 2017; Joshi et al., 2017; Kwiatkowski et al., 2019; Reddy et al., 2019; Jin et al., 2019; Wang et al., 2020), where the system is given a question and a particular (or a few) passage(s) and the answer must be found therein. In effect, MRC focuses on a particular core stage of a full QA pipeline, identifying answer spans, assuming that relevant documents and passages have already been identified. We also focus on this stage, adopting an MRC setting. Large generic (non domain-specific) MRC datasets have been constructed via crowd-annotation (Rajpurkar et al., 2016, 2018; Yang and Choi, 2019; Joshi et al., 2017), but crowd-annotation on that scale is difficult when biomedical expertise is required. An alternative is to *automatically* generate *cloze-style* MRC datasets. The last sentence or title of a random passage is treated as a question, some part (e.g., named entity) of the 'question' is masked, and the system is required to predict it. This approach has been used to generate large *artificial* cloze-style MRC datasets (Hill et al., 2016; Chen et al., 2016; Bajgar et al., 2016), including biomedical ones (Pappas et al., 2018, 2020). These datasets could be used to augment real ones, but have mostly been used as artificial experimental setups only.

When training examples for end-tasks are limited, as in realistic biomedical QA datasets, the currently dominant approach in NLP is to use pre-trained Transformers (Devlin et al., 2019; Liu et al., 2019; Lan et al., 2019; He et al., 2020; Raffel et al., 2020), possibly pre-trained on domain-specific

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 63–81
May 26, 2022.

corpora (Lee et al., 2019; Beltagy et al., 2019; Chalkidis et al., 2020), and fine-tune (further train) them on the limited examples of the end-tasks. Nevertheless, increasing the number of end-task examples typically improves performance. One way to achieve this is to employ *data augmentation* (DA) (Shorten et al., 2021; Feng et al., 2021), which adds artificial training instances to a training set, in our case the training set of the end task. It is unclear, however, which DA methods improve most (if at all) the performance of pre-trained models per end-task (Longpre et al., 2019, 2020). Consequently, Feng et al. (2021) recommend exploring when DA is effective for large pre-trained models.

In this paper, we thoroughly examine the impact of DA in biomedical QA, focusing on the factoid questions of the BIOASQ challenge (Tsatsaronis et al., 2015) in an MRC setting, i.e., we assume that relevant text passages, called *snippets* in BIOASQ, have already been identified. We first evaluate on BIOASQ three indicative off-the-shelf pre-trained models, DISTILBERT (Sanh et al., 2019), BIOBERT (Lee et al., 2019), ALBERT (Lan et al., 2019), already fine-tuned on SQUAD (Rajpurkar et al., 2016) or SQUAD-V2 (Rajpurkar et al., 2018), and we select ALBERT as our weak baseline. We also fine-tune ALBERT on BIOASQ, on top of its SQUAD fine-tuning, to obtain a stronger baseline. We then obtain additional artificial training instances from an artificial cloze-style MRC dataset, or via back-translation, information retrieval (IR), word substitution based on WORD2VEC or masked language modeling, question generation, or by extending the given passages with additional context. WORD2VEC-based word substitution, one of the simplest DA methods considered, improves test performance from 76.78% precision-recall AUC (for ALBERT fine-tuned on SQUAD and BIOASQ) to 84.99%. Although we focus on biomedical QA, our work should also be of interest in QA for other specialized domains, e.g., legal QA (Kien et al., 2020; Khazaeli et al., 2021; Zhang and Xing, 2021). Our work is the largest, in terms of DA methods considered, experimental study of DA for QA (Section 4).

Our main contributions are: (1) We present the largest (in terms of methods) experimental comparison of DA methods for QA, focusing on biomedical QA, where obtaining training instances is particularly difficult and costly. (2) We show that DA can lead to very large performance gains, even when using pre-trained Transformers fine-tuned on large generic (SQUAD) and/or small domain-specific (BIOASQ) end-task datasets, contributing to a broader discussion of if/when DA benefits pre-trained models. (3) We show that artificial cloze-style MRC datasets are useful for DA. (4) We show that one of the simplest DA methods, WORD2VEC-based word substitution, is also the best and is, therefore, recommended. (5) We make our artificial training examples and code publicly available.[1]

2 Experimental Setup

2.1 BIOASQ Data in a SQUAD setting

We experiment with data from BIOASQ-8 (2021), Phase B, Task B (Tsatsaronis et al., 2015), which contain English questions of biomedical experts. Each question is accompanied by (i) the gold answer (often several alternative phrasings) and (ii) gold relevant passages, called *snippets* (usually a single sentence each) from biomedical articles; the gold snippets contain the gold answer or other relevant information. There are four question types: *yes/no*, *factoid*, *list*, and questions requiring a *summary*. We focus on factoid questions (e.g., "*Which gene is involved in Giant Axonal Neuropathy?*").

We convert the BIOASQ data to triples each containing a question, a single gold snippet, and the span of the gold answer in the snippet, much as in SQUAD (Rajpurkar et al., 2016). If a question has multiple gold snippets, we produce equally many triples, discarding snippets that do not contain the gold answer. This conversion and considering only factoid questions allow us to use pre-trained Transformers already fine-tuned on SQUAD in a similar setting.[2] A disadvantage of the conversion is that our results are not directly comparable to those of BIOASQ. The goal of our work, however, is to study the effect of different DA methods on a modern Transformer-based QA baseline (and we show that fine-tuning it first on SQUAD helps), not to compete against BIOASQ participants, who often use models tailored to the particular competition.

From the 941 factoid questions of the original BIOASQ data, we obtained 3415 question-snippet-answer triples. We split these in training, development, test subsets (2848, 271, 296 triples, resp.), ensuring no question is in more than one subsets.

[1] See http://nlp.cs.aueb.gr/publications.html for links to the code and data.

[2] In the original BIOASQ data, multiple snippets may be given for a particular question, the answer may be present in several of them, and identifying any answer span suffices.

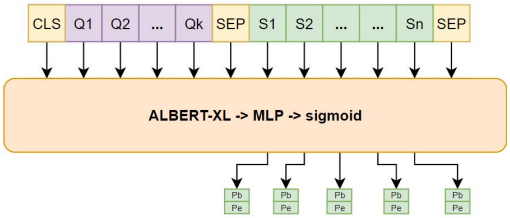

Figure 1: The model used in all of the following experiments. ALBERT-XL is fed with a question and snippet. Its contextualized embeddings are passed through an MLP with sigmoid activations that produces a begin (P_b) and end (P_e) probability per token of the snippet.

2.2 Off-the-shelf Models

As a starting point, we compared the performance of three publicly available pre-trained models that have already been fine-tuned for MRC on SQUAD (Rajpurkar et al., 2016) or SQUAD-V2 (Rajpurkar et al., 2018).[3] At the time of our experiments, ALBERT-based models (Lan et al., 2019) were among the best on the SQUAD leaderboards; here, we use ALBERT fine-tuned on SQUAD-V2. We also considered BIOBERT (Lee et al., 2019), because it is pre-trained on a biomedical corpus; again, we use it fine-tuned on SQUAD-V2. The third model, DISTILBERT (Sanh et al., 2019), was chosen because of its much smaller size, which makes running experiments easier. This model is pretrained on a generic corpus, like the original BERT, and we use it fine-tuned on SQUAD. All three models are used here off-the-self, i.e., they are only evaluated, not trained in any way on BIOASQ data. Throughout this work, we use the development subset of the BIOASQ data to select models and configurations of DA methods, but in this experiment we use the union of the training and development subsets, since no BIOASQ training is involved. ALBERT is the best off-the-shelf model considered (Table 1), hence we use it in all other experiments.[4]

Model	PRAUC (BIOASQ train+dev)
DISTILBERT (SQUAD)	64.27
BIOBERT (SQUAD-V2)	69.22
ALBERT (SQUAD-V2)	**75.05**

Table 1: *Off-the-shelf* pre-trained models, fine-tuned for MRC on SQUAD or SQUAD-V2. We report Precision-Recall AUC (PRAUC, %) on BIOASQ training and development data, since no BIOASQ training is involved.

[3]We obtained the models from https://huggingface.co/ktrapeznikov/albert-xlarge-v2-squad-v2. We use the XL version of ALBERT. The other two models adopt the BERT-BASE architecture; no XL variants were available.

[4]We discuss PRAUC in Sections 2.3 and 2.4.

2.3 Model Architecture Modifications

The results of Table 1 were obtained by feeding the three off-the-shelf models with the concatenation of the question and snippet of each question-snippet-answer BIOASQ triple (training or development), without training of any kind. Following a typical MRC architecture, each model was previously fine-tuned (by others) on SQUAD (or SQUAD-V2) with a shared dense layer on top of each contextualized token embedding (of the snippet only) that the pre-trained model generates. The dense layer produces two logits per token, indicating the model's confidence that the token is the beginning or end of the answer, respectively. Two separate softmax activations operate across all the begin and end logits, respectively, and the answer is the span (of the snippet) whose first and last tokens have the highest sum of begin and end probabilities (and the correct order).[5] The two softmax activations presuppose that there is a single contiguous answer span in each snippet. This is true in SQUAD, but in our BIOASQ data the (single) answer of a triple may consist of multiple non-contiguous spans of the triple's snippet (this happens in 583 out of 2,848 training instances). Hence, in the following experiments, where we further fine-tune ALBERT on BIOASQ or artificial training data, we replace the two softmax activations by two sigmoids that produce the begin and end probability per token of the snippet. Any token whose begin (or end) probability is above a threshold t is treated as the beginning (or end) of an answer span. The PRAUC evaluation measure (discussed below) aggregates results over different t values. We also replace the dense layer on top of the contextualized token embeddings by a Multi-Layer Perceptron (MLP) with a single hidden layer, which performed better on our development data. We use this single typical MRC model architecture (Fig.1) in all the following experiments, since we aim to study the effect of several DA methods, not to propose new MRC model architectures.

2.4 Evaluation Measure

Given a development or test question-snippet-answer triple and a decision threshold t (Section 2.3), we compute precision and recall at the token level, i.e., we measure the ability of the model to identify the tokens of the answer. Precision is the number of correctly identified answer tokens,

[5]In SQUAD-V2, additional layers decide if a question is answerable. We do not discuss them to save space.

divided by the number of tokens in the model's answer. Recall is the number of correctly identified answer tokens, divided by the number of tokens in the correct answer. For different thresholds t, we obtain different precision-recall pairs for the same question-snippet-answer triple, which can be plotted as a precision-recall curve. We compute the trapezoidal area under the precision-recall curve (PRAUC), and we then macro-average the PRAUC scores over the test (or development) triples.[6]

2.5 Baselines

We use two baselines that do not involve DA: i) off-the-shelf ALBERT, pre-trained on a generic corpus, already fine-tuned on SQUAD-V2 (last model of Table 1); and ii) same as the first baseline, but further fine-tuned (on top of the fine-tuning on SQUAD-V2) on our BIOASQ training data, with the modified architecture of Section 2.3. Table 2 shows that the second baseline is much stronger. Hence, we report performance gains with DA methods against the second baseline in subsequent sections.[7]

Model	+train ex.	PRAUC (BIOASQ dev)
ALBERT (SQUAD-V2)	0	80.25
+BIOASQ	2,848	89.57

Table 2: Performance of baselines on BIOASQ dev. data. The first one is ALBERT-XL fine-tuned on SQUAD-V2. The second one is also fine-tuned on BIOASQ, with the modified architecture of Fig. 1. We also show the number of domain-specific (BIOASQ) training examples.

3 Data Augmentation Methods

There are two alternatives when using the artificial training instances that DA generates (Yang et al., 2019). In our case, we always start with ALBERT, pre-trained on a generic corpus, and already fine-tuned on SQUAD-V2. In the first alternative, the model is then further fine-tuned on the artificial instances, and is then finally fine-tuned on the end-task data (BIOASQ). In the second alternative, the artificial and the end-task data are mixed, and the model is fine-tuned on the mixed data. In each experiment below, we use the alternative (among the two) that leads to the best development PRAUC.

3.1 Artificial Cloze-style MRC Dataset

For this augmentation method, we use BIOMRC (Pappas et al., 2020), the most recent and largest *artificial* cloze-style biomedical MRC dataset. BIOMRC comes in two versions, LARGE and LITE, with 813k and 100k cloze-style questions, respectively. We use BIOMRC LITE. Each 'question' is the title of a biomedical article, with an entity mentioned in the title hidden. Each question is accompanied by a passage (the abstract of the article), candidate answers (entities mentioned in the abstract), and the gold answer. From each passage we keep only the sentence containing the gold answer as the given snippet, and we generate a question-snippet-answer triple.[8] If more than one sentences of the passage contain the gold answer, we create multiple triples, one for each sentence. We end up with approximately 142k artificial training triples.

ALBERT (SQUAD-V2)	+train ex.	PRAUC (BIOASQ dev)
+BIOASQ	2,848	89.57
+BIOMRC	2,848	78.66
+BIOMRC +BIOASQ	5,696	91.57
+BIOMRC	10,000	91.20
+BIOMRC +BIOASQ	12,848	**93.15**
+BIOMRC	30,000	90.57
+BIOMRC +BIOASQ	32,848	92.19
+BIOMRC	50,000	91.19
+BIOMRC +BIOASQ	52,848	91.51
+BIOMRC	100,000	90.93
+BIOMRC +BIOASQ	102,848	92.39

Table 3: Adding training examples from an *artificial cloze-style* MRC dataset (BIOMRC). The '+train ex.' column shows the number of domain-specific training examples (from BIOMRC and/or BIOASQ) used, on top of examples seen during fine-tuning on SQUAD-V2.

In Table 3, the starting point is the weak baseline of Table 2 (ALBERT fine-tuned on SQUAD-V2). We compare to the strong baseline (the second one of Table 2), which was further fine-tuned on BIOASQ (+BIOASQ). We show results when fine-tuning on BIOMRC (+BIOMRC) instead of BIOASQ, and when fine-tuning on both BIOMRC and BIOASQ (+BIOMRC +BIOASQ), using 10k to 100k randomly sampled BIOMRC examples. Interestingly, fine-tuning on 10k artificial BIOMRC examples leads to better performance (91.20 dev. PRAUC) than fine-tuning on BIOASQ (89.57). The best performance (93.15) is obtained by fine-tuning on both BIOASQ and 10k BIOMRC examples. We attribute this improvement to the resemblance of BIOMRC to BIOASQ data. We see no benefit when adding more than 10k BIOMRC examples, which may be an indication that the useful (for BIOASQ) patterns that the model can learn from BIOMRC are limited.

[6] PRAUC is similar to Mean Average Precision (Manning et al., 2008), but obtains precision-recall points differently.

[7] We also experimented pre-trained ALBERT directly fine-tuned only on BIOASQ, but the performance was much worse.

[8] See the appendix for examples of all the DA methods.

3.2 Back-translation

Back translation (BTR) has been used for data augmentation in several NLP tasks (Feng et al., 2021; Shorten et al., 2021). The training examples are machine-translated from a source to a pivot language and back, obtaining paraphrases. We initially used French as the pivot language, then also Spanish and German, always translating from English to a pivot language and back with Google Translate. For each question-snippet-answer training triple of BIOASQ, we generate two new triples by back-translating either the question or the snippet. If a new triple is identical to the original one, we discard it. We obtained 4,901 new training examples pivoting only to French, and 15,593 when also pivoting to Spanish and German.

ALBERT (SQUAD-V2)	+train ex.	PRAUC (BIOASQ dev)
+BIOASQ	2,848	89.57
+BTR [FR]	2,848	91.84
+BTR [FR] +BIOASQ	5,696	**92.95**
+BTR [FR]	4,901	89.80
+BTR [FR] +BIOASQ	7,749	91.44
+BTR [FR,ES,DE]	2,848	89.80
+BTR [FR,ES,DE] +BIOASQ	5,696	89.99
+BTR [FR,ES,DE]	14,229	92.21
+BTR [FR,ES,DE] +BIOASQ	17,077	92.21

Table 4: Data augmentation via *back-translation* (BTR), using one (FR) or three (FR, ES, DE) pivot languages.

Table 3 shows that adding back-translations to the BIOASQ training data increases development PRAUC from 89.57 to 91.44 (or 92.66) with one (or three) pivot languages. Using back-translations with one pivot (+BTR [FR]) instead of the original BIOASQ data slightly surpasses the strong baseline (89.80 vs. 89.57); and with three pivots, using only back-translations (+BTR [FR,DE,ES]) performs almost the same as adding the original BIOASQ data too (92.52 vs. 92.66). These results show that BTR produces very good training instances and that further benefits may be possible with more pivots. Nevertheless simpler methods (e.g., WORD2VEC-based word substitution, discussed below) offer larger gains with fewer artificial training instances.

3.3 Information Retrieval

Data augmentation based on Information Retrieval (IR) has been found promising in previous open-domain QA work (Yang et al., 2019).[9] Given a question and a gold answer, the question is used as a query to an IR system. Any retrieved document (or passage therein) that includes the gold answer is used to construct a new training example (with the same question and gold answer). We used the open data from the PUBMED Baseline Repository[10] to create a pool of 22.3M biomedical documents. Each document is the concatenation of the title and abstract of a PUBMED article. We indexed all documents with an ElasticSearch retrieval engine[11] and used the 500 top ranked (by BM25) documents per question. From the original 2,848 question-snippet-answer triples, only 289 more were generated, because in most of the retrieved documents no sentence included the entire answer (individual terms of the answer might be scattered in the document). We suspect that the biomedical experts of BIOASQ create questions whose answers cannot be found in large numbers of documents (unlike common questions in open-domain QA datasets), and the few relevant documents (and snippets) of each question have already been included in the BIOASQ training data. Table 5 shows that IR-based augmentation led to very minor gains in our case, because of the very few additional instances.

ALBERT (SQUAD-V2)	+train ex.	PRAUC (BIOASQ dev)
+BIOASQ	2,848	89.57
+IR	289	80.30
+IR +BIOASQ	3,137	**89.80**

Table 5: Data augmentation via *information retrieval* (IR), using PUBMED titles and abstracts as documents.

3.4 Word Substitution

These methods replace words of the original training examples by similar words (e.g., synonyms) from a thesaurus (Jungiewicz and Smywinski-Pohl, 2019; Abdollahi et al., 2020) or words with similar embeddings (Wang and Yang, 2015). More recent work uses large language models, pre-trained to predict masked tokens, which suggest replacements of randomly masked words of the original examples (Kobayashi, 2018; Wu et al., 2019).

3.4.1 WORD2VEC-based Word Substitution

In this case, we use biomedical WORD2VEC (Mikolov et al., 2013; Brokos et al., 2018) embeddings. Given a question-snippet-answer training instance, we consider all the word tokens of the snippet (excluding stop-words). For each token w_i ($i = 1, \ldots, n$) of the snippet, we select the $k_i \leq K$ most similar words w_j ($j = 1, \ldots, k_i$) of

[9] Yang et al. (2019) gained 2.7 to 9.7 F1 percentage points (pp.) in all four datasets they experimented with.

[10] lhncbc.nlm.nih.gov/ii/information/MBR.html

[11] https://www.elastic.co/

the vocabulary, using cosine similarity of word embeddings $(\vec{w}_i, \vec{w}_j)$, that satisfy $\cos(\vec{w}_i, \vec{w}_j) \geq C$. We then produce $(k_1+1)(k_2+1)\ldots(k_n+1)-1$ artificial training instances by replacing each token w_i of the snippet by one of its k_i most similar words (or itself), requiring at least one token of the original snippet to have been replaced. We then randomly sample 10k to 100k of the resulting instances and use them as additional training examples. We set $K = 10$, $C = 0.95$ based on preliminary experiments on development data. Adding 10k of the resulting artificial training examples to the original BIOASQ examples leads to 95.60 development PRAUC, outperforming the strong baseline (89.57) by six percentage points (Table 6). Using only the 10k artificial examples, without any of the original examples, achieves almost identical performance (95.59), which suggests that the generated examples are of high quality. As when using artificial MRC examples (Table 3), adding more than 10k artificial instances provides no further benefit, probably because we end up adding too many minor variants of the same original examples.

ALBERT (SQUAD-V2)	+train ex.	PRAUC (BIOASQ dev)
+ BIOASQ	2,848	89.57
+WORD2VEC	2,848	95.56
+WORD2VEC +BIOASQ	5,696	95.27
+WORD2VEC	10,000	95.59
+WORD2VEC +BIOASQ	12,848	**95.60**
+WORD2VEC	30,000	95.28
+WORD2VEC +BIOASQ	32,848	95.20
+WORD2VEC	50,000	95.16
+WORD2VEC +BIOASQ	52,848	95.13
+WORD2VEC	100,000	95.36
+WORD2VEC +BIOASQ	102,848	95.22

Table 6: Data augmentation with WORD2VEC-*based word substitution*, using biomedical embeddings.

The same DA mechanism could have been applied to questions instead of snippets. In preliminary experiments, we employed an additional pre-trained natural language inference (NLI) model (El Boukkouri et al., 2020) as a *consistency* mechanism to ensure the modified snippets followed from the original ones, but this also greatly reduced the number of artificial training instances we could generate. Performance was better without this mechanism, i.e., generating more artificial instances was better than generating fewer higher quality ones.

3.4.2 Masked LM Word Substitution

Here we use BIOLM (Lewis et al., 2020) and specifically a ROBERTA-LARGE model pre-trained on PUBMED, PMC, and MIMIC-III (Zhu et al., 2018) with a new vocabulary extracted from PUBMED.[12] We use the same process as in WORD2VEC word substitution, but each candidate replacement w_j of an original word w_i of the snippet must now satisfy $p(w_j) \geq P$ (instead of $\cos(\vec{w}_i, \vec{w}_j) \geq C$), where $p(w_j)$ is the probability assigned to w_j by the pre-trained model; we also rank the candidate replacements w_j of each w_i by $p(w_j)$. We set $P = 0.95$, based on preliminary experiments on development data. Table 7 shows that BIOLM-based substitution is almost as good as WORD2VEC-based substitution (94.45 vs. 95.60), but for BIOLM the best performance is obtained with 50k artificial examples (compared to 10k for WORD2VEC). This is probably due to the fact that BIOLM suggests words that fit the particular context of the word being replaced and may, thus, suggest words with very different meanings that can be used in the particular context, adding noisy examples. By contrast, when using WORD2VEC we compare more directly each original word with candidate replacements.[13]

ALBERT (SQUAD-V2)	+train ex.	PRAUC (BIOASQ dev)
+BIOASQ	2,848	89.57
+BIOLM	2,848	91.76
+BIOLM +BIOASQ	5,696	92.37
+BIOLM	10,000	94.06
+BIOLM +BIOASQ	12,848	94.06
+BIOLM	30,000	93.63
+BIOLM +BIOASQ	32,848	93.75
+BIOLM	50,000	93.94
+BIOLM +BIOASQ	52,848	**94.45**
+BIOLM	100,000	93.79
+BIOLM +BIOASQ	102,848	93.84

Table 7: Data augmentation with *word substitution* based on *masked language modeling* using BIOLM.

3.5 Question Generation

Question generation (QG) has been found an effective DA method in open-domain MRC (Alberti et al., 2019; Chan and Fan, 2019; Lopez et al., 2020). The main reported benefit is that it increases the diversity of questions (Qiu and Xiong, 2019; Sultan et al., 2020). Typically QG models are fed with a snippet s, select an answer span a of s, and generate a question q answered by a. We take T5 (Raffel et al., 2020) fine-tuned for QG on a modified version of SQUAD by Lopez et al. (2020)[14] and use it to gen-

[12] We did not use BIOLM as an off-the-shelf QA model (Section 2.2), because it was not available fine-tuned on SQUAD.

[13] WORD2VEC embeddings are not sensitive to the particular context of the snippet and rely exclusively on the (many more) contexts of each word encountered in the pre-training corpus.

[14] The T5 QG model we used is available at https://github.com/patil-suraj/question_generation.

erate alternative questions q' and answer spans a' from the snippets s of the BIOASQ $\langle q, s, a\rangle$ training triples, producing artificial $\langle q', s, a'\rangle$ triples. Multiple artificial triples can be generated from the same original one (the same s), but we require each q' to be answered by a different answer span a' to maximize the diversity of questions. We obtained 3,389 artificial triples from the 2,848 original ones this way. An alternative we explored is to select random snippets s from random PUBMED abstracts, and use the QG model to produce artificial $\langle q', s, a'\rangle$ triples. The alternative approach can generate millions of artificial triples; we generated up to 100k.

ALBERT (SQUAD-V2)	+train ex.	PRAUC (BIOASQ dev)
+BIOASQ	2,848	89.57
+T5@BIOASQ	3,389	84.46
+T5@BIOASQ +BIOASQ	6,237	88.46
+T5@PUBMED	2,848	85.79
+T5@PUBMED +BIOASQ	5,696	89.29
+T5@PUBMED	10,000	87.30
+T5@PUBMED +BIOASQ	12,848	89.34
+T5@PUBMED	30,000	86.65
+T5@PUBMED +BIOASQ	32,848	90.51
+T5@PUBMED	50,000	87.30
+T5@PUBMED +BIOASQ	52,848	**90.69**
+T5@PUBMED	100,000	87.30
+T5@PUBMED +BIOASQ	102,848	90.61

Table 8: Data augmentation via *question generation* using T5. Questions are generated from the training snippets of BIOASQ (T5@BIOASQ) or from random snippets from random PUBMED abstracts (T5@PUBMED).

Table 8 shows that adding to the BIOASQ training data the artificial triples we obtained from BIOASQ (+T5@BIOASQ, BIOASQ) is worse (88.46 vs. 89.57) than our strong baseline (+BIOASQ). Fine-tuning only on the artificial triples (+T5@BIOASQ) is much worse (84.46), i.e., the artificial triples are much less useful, despite being more than the original ones. Adding artificial triples from PUBMED (+T5@PUBMED, BIOASQ) performs only slightly better (90.69) than the strong baseline, when using 50k artificial triples, with no further benefit when using more. A possible explanation for these poor results is the T5was fine-tuned for QG on the open-domain SQUAD dataset. Thus, the generated questions are rather simplistic and not indicative of the specialized questions of BIOASQ. Indeed, most of the generated questions are minor rephrases of the given snippet (e.g., subject replaced by 'what').

3.6 Adding Context

In the original training question-snippet-answer $\langle q, s, a\rangle$ triples, s is usually a single sentence. To help the QA model learn to better distinguish relevant from irrelevant parts of the given snippet, we experimented with an additional DA method, where we find the original article that s comes from and we expand s with the k_1 (and k_2) sentences preceding (and following) it.[15] For each original $\langle q, s, a\rangle$ triple, we create multiple new $\langle q, s', a\rangle$ artificial triples, for different values of $k_1 \geq 0$ and $k_2 \geq 0$, such that $k_1 + k_2 = K$.[16] We experiment with $K = 2$ (three new triples for each original one); then to obtain more artificial examples, we repeat with $K = 4$ (five new triples for each original). To avoid truncation of the input examples, we remove all artificial examples that exceed 500 characters in length. For $K \in \{2, 4\}$, we obtain a development PRAUC score of 94.21 (Table 9), which is surpassed only by the the two embedding-based word substitution methods (Tables 6–7). This DA method was introduced by Yoon et al. (2020), who used it in BIOASQ.[17]

ALBERT (SQUAD-V2)	+train ex.	PRAUC (BIOASQ dev)
+BIOASQ	2,848	89.57
+CONTEXT ($K = 2$)	4,568	93.91
+CONTEXT ($K = 2$) +BIOASQ	7,416	94.05
+CONTEXT ($K \in \{2, 4\}$)	6,428	94.20
+CONTEXT ($K \in \{2, 4\}$) +BIOASQ	9,276	**94.21**

Table 9: Data augmentation by *adding context to the snippet* ($K = 2$ or $K \in \{2, 4\}$ surrounding sentences).

3.7 Final Results

Table 10 shows the performance of all the DA methods considered, on both development and test data. For each DA method, we use the configuration (from Tables 3–9) with the best development score. The test scores are lower than the corresponding development ones, since several hyper-parameters (e.g., K, C in the case of WORD2VEC-based word substitution, number of training epochs) are tuned on the development set. The test set also seems to be harder than the development one, since our weak baseline (ALBERT fine-tuned on SQUAD-V2 with no further training) also performs worse on the test set (77.78 vs. 80.25). Nevertheless, the test scores confirm that WORD2VEC-based word substitution is the best DA method considered, leading to a performance gain of 8.2 percentage points test PRAUC compared to the strong baseline (84.99 vs. 77.78). The ranking of the other DA methods

[15] In BIOASQ, each gold snippet is accompanied by the PUBMED id of the article it was extracted from.

[16] Simply setting $k_1 = k_2$ would risk misguiding the model to always prefer the central sentence. We also experimented with *random* k_1 (or k_2) sentences before (and after) s, but performance was much worse, possibly because the random sentences led to inferior context-aware token embeddings.

[17] Yoon et al. (2020) reported an improvement in BIOASQ's Lenient Accuracy by 2.49 percentage points.

does not change when ranking by test score, instead of development score, with the only exception of adding context to the given passage (+CONTEXT), which is now slightly worse than adding instances from the artificial BIOMRC dataset. Interestingly, all the DA methods, even the weakest IR-based one, improve upon the test score of the strong baseline.

Method	+train ex.	PRAUC (dev)	PRAUC (test)
ALBERT (SQUAD-V2)	0	80.25	77.78
+ BIOASQ	2,848	89.57	76.78
+WORD2VEC +BIOASQ	12,848	**95.60 (+6.03)**	**84.99 (+8.21)**
+BIOLM +BIOASQ	52,848	94.45 (+4.88)	82.76 (+5.98)
+CONTEXT +BIOASQ	9,276	94.21 (+4.64)	81.63 (+4.85)
+BIOMRC +BIOASQ	12,848	93.15 (+3.58)	82.04 (+5.26)
+BTR +BIOASQ	18,441	92.66 (+3.09)	81.27 (+4.49)
+T5@PUBMED +BIOASQ	52,848	90.69 (+1.12)	80.26 (+3.48)
+IR +BIOASQ	3,137	89.80 (+0.23)	78.66 (+1.88)

Table 10: Performance of DA methods on *development* and *test* data, ordered by decreasing development score. For each DA method, we use the configuration (from Tables 3–9) with the best development score.

4 Related Work

DA is a key ingredient of success in deep learning for computer vision (Shorten and Khoshgoftaar, 2019). DA for NLP has been explored less, but is an active research area (Shorten et al., 2021; Feng et al., 2021), with methods ranging from leveraging knowledge graphs (Moussallem et al., 2019) to generating textual data from scratch (Yang et al., 2020; Bayer et al., 2021a). The most common NLP task in DA is text classification (Bayer et al., 2021b). Feng et al. (2021) consider span-based NLP tasks in specialized domains, which includes biomedical MRC, among the most challenging for DA.

Word substitution is a simple and common DA approach in NLP. In thesaurus-based substitution (Jungiewicz and Smywinski-Pohl, 2019; Abdollahi et al., 2020), words are replaced by synonyms or closely related words (e.g., hypernyms). Word embedding substitution (Wang and Yang, 2015) replaces words by others nearby in a pre-trained vector space model (Section 3.4). Alternatively, a random word is removed, inserted (Wei and Zou, 2019a; Miao et al., 2020), or noised with spelling errors (Belinkov and Bisk, 2018). Sentences may also be re-ordered or removed (Shen et al., 2020; Chen et al., 2021). Text generation has also been used in several NLP tasks for adversarial augmentation (Cheng et al., 2020), to paraphrase training examples (Ribeiro et al., 2018; Cai et al., 2020; Xie et al., 2020), or generate new (Anaby-Tavor et al., 2020; Kumar et al., 2020). Back-translation (Sennrich et al., 2016) is also widely used across NLP tasks (Shorten et al., 2021; Feng et al., 2021).

DA work for QA in particular includes back-translation (Du et al., 2019), question generation (Zhang and Bansal, 2019; Alberti et al., 2019; Chan and Fan, 2019; Lopez et al., 2020; Yang et al., 2020), paraphrasing (Dong et al., 2017; Liu et al., 2020), and synonym replacement (Nugraha and Suyanto, 2019), but not in a biomedical setting. The IR-based DA we used (Section 3.3) follows Yang et al. (2019), who experimented in English and Chinese, but not in the biomedical domain. Expanding the passage with surrounding sentences (Section 3.6) follows Yoon et al. (2020), who used this method in BIOASQ. Dhingra et al. (2018) create artificial cloze-style MRC datasets and use them to pre-train neural QA models (not Transformers), which are then fine-tuned on real training examples. By contrast, we use artificial MRC datasets to fine-tune large pre-trained Transformers. All the above studies experimentally compare at most two DA methods; we compare seven. Hence, our work is the largest (in terms of methods considered) experimental study of DA for QA (and possibly NLP).

Longpre et al. (2019) report that back-translation did not improve generalization in (non-biomedical) QA experiments with fine-tuned pre-trained Transformers. Longpre et al. (2020) report that back-translation and Easy Data Augmentation (Wei and Zou, 2019b) are not always effective in text classification when fine-tuning pretrained Transformers, even with small end-task training sets. Consequently, Feng et al. (2021) recommend exploring when DA is effective for large pre-trained models. Our work contributes in this discussion by showing that DA can lead to very significant performance gains, even when using large pre-trained Transformers fine-tuned on large generic (SQUAD) and/or small domain-specific (BIOASQ) end-task datasets.

5 Limitations and Future Work

A limitation of our work is that we consider only DA in the *input space*, i.e., the artificial instances are in textual form, like the original ones, as opposed to, e.g., interpolating feature vectors (Chawla et al., 2002; DeVries and Taylor, 2017; Shorten et al., 2021). We also consider only *offline* augmentation, i.e., the artificial instances are generated once, before training, as opposed to artificial instances generated anew for each training epoch. These two limitations, which are common in DA for NLP, allow generating model-agnostic training instances

once and reusing them across training epochs and different models. This greatly reduces computation costs and allows sharing the augmented datasets. Online DA, however, exposes the model to many more synthetic data; and feature space DA may act as layer-specific regularization. One could also exploit ideas from active learning (Ein-Dor et al., 2020; Margatina et al., 2021) to select the most informative, diverse, and representative artificial training instances among those that DA generates. Small subsets of the BIOASQ data could also be used to study the effect of DA in few-shot learning.

References

Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, and Jinyan Li. 2020. Ontology-guided data augmentation for medical document classification. In *Artificial Intelligence in Medicine*, pages 78–88, Cham. Springer International Publishing.

Ahmad Aghaebrahimian. 2018. Linguistically-based deep unstructured question answering. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 433–443, Brussels, Belgium. Association for Computational Linguistics.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7383–7390.

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2016. Embracing data abundance: BookTest Dataset for Reading Comprehension. *CoRR*.

Markus Bayer, Marc-André Kaufhold, Björn Buchhold, Marcel Keller, J. Dallmeyer, and Christian Reuter. 2021a. Data augmentation in natural language processing: A novel text generation approach for long and short text classifiers. *ArXiv*, abs/2103.14453.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2021b. A survey on data augmentation for text classification. *CoRR*, abs/2107.03158.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. *ArXiv*, abs/1711.02173.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

George Brokos, Polyvios Liosis, Ryan McDonald, Dimitris Pappas, and Ion Androutsopoulos. 2018. AUEB at BioASQ 6: Document and Snippet Retrieval. In *Proceedings of the 6th BioASQ Workshop*, pages 30–39, Brussels, Belgium.

Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin. 2020. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6334–6343, Online. Association for Computational Linguistics.

Daniel Fernando Campos, T. Nguyen, M. Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, L. Deng, and Bhaskar Mitra. 2016. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, abs/1611.09268.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162, Hong Kong, China. Association for Computational Linguistics.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321—-357.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017a. Reading wikipedia to answer open-domain questions. *CoRR*, abs/1704.00051.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017b. Reading Wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*.

Jiaao Chen, Dinghan Shen, Weizhu Chen, and Diyi Yang. 2021. Hiddencut: Simple data augmentation for natural language understanding with better generalization. *ArXiv*, abs/2106.00149.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. Advaug: Robust adversarial augmentation for neural machine translation. *ArXiv*, abs/2006.11834.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Terrance DeVries and Graham W. Taylor. 2017. Dataset augmentation in feature space.

Bhuwan Dhingra, Danish Danish, and Dheeraj Rajagopal. 2018. Simple and effective semi-supervised question answering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 582–587, New Orleans, Louisiana.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark. Association for Computational Linguistics.

Y. Du, W. Guo, and Y. Zhao. 2019. Hierarchical question-aware context learning with augmented data for biomedical question answering. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 370–375.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online.

Bin Fu, Yunqi Qiu, Chengguang Tang, Y. Li, H. Yu, and J. Sun. 2020. A survey on complex question answering over knowledge base: Recent advances and challenges. *ArXiv*, abs/2007.13069.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *CoRR*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Michał Jungiewicz and Aleksander Smywinski-Pohl. 2019. Towards textual data augmentation for neural networks: synonyms and maximum loss. *Computer Science*, 20.

Soha Khazaeli, Janardhana Punuru, Chad Morris, Sanjay Sharma, Bert Staub, Michael Cole, Sunny Chiu-Webster, and Dhruv Sakalley. 2021. A free format legal question answering system. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 107–113, Punta Cana, Dominican Republic.

Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. 2020. Answering legal questions by learning neural attentive text representation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 988–998, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. 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. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Dayiheng Liu, Yeyun Gong, Jie Fu, Yu Yan, Jiusheng Chen, Jiancheng Lv, Nan Duan, and Ming Zhou. 2020. Tell me how to ask again: Question data augmentation with controllable rewriting in continuous space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5798–5810, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Shayne Longpre, Yi Lu, Zhucheng Tu, and Chris DuBois. 2019. An exploration of data augmentation and sampling techniques for domain-agnostic question answering. *CoRR*, abs/1912.02145.

Shayne Longpre, Yu Wang, and Chris DuBois. 2020. How effective is task-agnostic data augmentation for pretrained transformers? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4401–4411, Online.

Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, and Charibeth Cheng. 2020. Transformer-based end-to-end question generation. *CoRR*, abs/2005.01107.

Kangqi Luo, Fengli Lin, Xusheng Luo, and Kenny Zhu. 2018. Knowledge base question answering via encoding of complex query graphs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2185–2194, Brussels, Belgium. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic.

Zhengjie Miao, Yuliang Li, Xiaolan Wang, and Wang-Chiew Tan. 2020. Snippext: Semi-supervised opinion mining with augmented data. In *Proceedings of The Web Conference 2020*, page 617–628, New York, NY, USA. Association for Computing Machinery.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Diego Moussallem, Mihael Arcan, Axel-Cyrille Ngonga Ngomo, and Paul Buitelaar. 2019. Augmenting neural machine translation with knowledge graphs. *CoRR*, abs/1902.08816.

H. S. Nugraha and S. Suyanto. 2019. Typographic-based data augmentation to improve a question retrieval in short dialogue system. In *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 44–49.

Dimitris Pappas and Ion Androutsopoulos. 2021. A neural model for joint document and snippet ranking in question answering for large document collections. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*

and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3896–3907, Online. Association for Computational Linguistics.

Dimitris Pappas, Ion Androutsopoulos, and Haris Papageorgiou. 2018. BioRead: A new dataset for biomedical reading comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. 2020. BioMRC: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 140–149, Online. Association for Computational Linguistics.

Jiazuo Qiu and Deyi Xiong. 2019. Generating highly relevant questions. *CoRR*, abs/1910.03401.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.

Dinghan Shen, Ming Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *ArXiv*, abs/2009.13818.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(60).

Connor Shorten, Taghi M. Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of Big Data*, 8(101).

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656, Online. Association for Computational Linguistics.

G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M.R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artieres, A. Ngonga, N. Heino, E. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, and G. Paliouras. 2015. An Overview of the BioASQ Large-Scale Biomedical Semantic Indexing and Question Answering Competition. *BMC Bioinformatics*, 16(138).

Bingning Wang, Ting Yao, Qi Zhang, Jingfang Xu, and Xiaochuan Wang. 2020. Reco: A large scale chinese reading comprehension dataset on opinion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9146–9153.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019a. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019b. EDA: Easy data augmentation techniques for boosting performance

on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *International Conference on Computational Science 2019*, pages 84–95, Cham.

Qizhe Xie, Zihang Dai, E. Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training. *arXiv: Learning*.

Naganand Yadati, Dayanidhi R S, Vaishnavi S, Indira K M, and Srinidhi G. 2021. Knowledge base question answering through recursive hypergraphs. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 448–454, Online. Association for Computational Linguistics.

Wei Yang, Yuqing Xie, Luchen Tan, Kun Xiong, M. Li, and Jimmy Lin. 2019. Data augmentation for bert fine-tuning in open-domain question answering. *ArXiv*, abs/1904.06652.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025, Online. Association for Computational Linguistics.

Zhengzhe Yang and Jinho D. Choi. 2019. FriendsQA: Open-domain question answering on TV show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197, Stockholm, Sweden. Association for Computational Linguistics.

Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. 2020. Pre-trained language model for biomedical question answering. In *Machine Learning and Knowledge Discovery in Databases*, pages 727–740, Cham. Springer International Publishing.

Na-Na Zhang and Yinan Xing. 2021. Questions and answers on legal texts based on BERT-BiGRU. *Journal of Physics: Conference Series*, 1828(1):012035.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2495–2509, Hong Kong, China. Association for Computational Linguistics.

Henghui Zhu, Ioannis Ch Paschalidis, and Amir Tahmasebi. 2018. Clinical concept extraction with contextual word embedding. *arXiv preprint arXiv:1810.10566*.

Appendix

A Combining Augmentation Methods

We also tried to combine DA methods. In Table A1, we incrementally add to the training set of the strong baseline (ALBERT fine-tuned on SQUAD-V2, then BIOASQ) artificial training examples obtained from WORD2VEC-based word substitution, then (additionally) training examples obtained by expanding the context of the given passage etc. We started with the artificial examples of the WORD2VEC-based method, which had the best development score, skipped the other (BIOLM-based) word substitution method, then continued with examples from BIOMRC and back-translation, which were the next best in terms of development score. Unfortunately, there was no significant gain, compared to using only the WORD2VEC-based method, which suggests that the DA methods we consider are not complementary. An alternative approach would be to stack DA methods, instead of accumulating their training examples. For example, one could apply the WORD2VEC method to artificial examples produced by increasing the context of the given passages. We leave this for future work.

Method	+train ex.	PRAUC (dev)	PRAUC (test)
ALBERT (SQUAD-V2)	0	80.25	77.78
+BIOASQ	2,848	89.57	76.78
+ WORD2VEC	12,848	**95.60**	84.99
+ CONTEXT	19,276	93.98	83.54
+ BIOMRC	29,276	94.27	85.18
+ BTR	44,869	93.44	83.97

Table A1: Results using a combination of Context Increasing and WORD2VEC data augmentation.

B Examples of Artificial Data

B.1 BIOMRC

Table D3 presents training instances generated from the BIOMRC dataset. Each instance is a triple containing a cloze-style question, a snippet, and the span of the snippet answering a question. This is very similar to the SQUAD setting which we have adopted in our experiments (see Section 2.1).

B.2 Back-translation

Tables D4 and D5 show training instances generated via back-translation of BIOASQ questions and snippets, respectively. The back-translated questions and snippets retain the semantics of the original ones while adding diversity to the training set.

B.3 Information Retrieval

Table D6 contains training instances generated via Information Retrieval. A BIOASQ question is used as a query in a search engine to retrieve PUBMED documents (abstracts and titles). From the retrieved documents all the snippets containing the answer are extracted and used to generate new training triples. Note that although a retrieved snippet may contain the entity that answers the BIOASQ question, it is not always evident that it answers the question, e.g., it may answer another question as is the case in the instance with id 29767248.

B.4 Word Substitution

Tables D7 and D8 presents examples generated via word substitution based on WORD2VEC and BIOLM respectively. Although some substitutions may induce noise, the generated snippets tend to retain the semantics of the original ones and add diversity to the training set.

B.5 Question Generation

Tables D9 and D10 show examples generated via Question Generation using BIOASQ snippets and random snippets from random PUBMED articles respectively. Although, the generated triples introduce diverse answers they contain rather simplistic questions which are not indicative of the specialized questions found in BIOASQ.

B.6 Additional Context

Table D11 contains examples generated by adding context to the original BIOASQ snippets. The additional context provides additional information that helps the model to better distinguish relevant and irrelevant parts of the original snippet.

C Computing Infrastructure

All of our experiments run on a titan-X GPU with 12GB of Memory while all code was compiled for CUDA Version 10.2. The personal computer we used offers 32GB of DDR4-RAM Memory and a 6-core Intel(R) Core(TM) i7-5820K CPU.

D Hyper-parameter tuning

The random seed in all experiments was set to 1. For data augmentation through Information Retrieval (IR), we use an ElasticSearch cluster to retrieve relevant abstracts using BM25 with default parameters.

Due to computational and time restrictions, hyper-parameter tuning was performed with grid-search by training on the original 2,848 BIOASQ examples (Table 2), i.e., without data augmentation, and evaluating on the development data. The 'best' hyper-parameter values were then used in all the augmentation experiments. The hyper-parameter search space (48 settings) and the selected values can be seen in Table D2.

Hyperparameter	choices	best dev. setting
Random Seed	{1}	1
MLP Hidden Size	{50, 100}	100
Total Epochs	{50, 100}	50
Patience	{5}	5
Monitor Score	{AUC, loss}	AUC
Learning Rate	{0.1, 0.01, 2e-5, 5e-5 }	5e-5
Weight Decay	{0.01}	0.01
Warmup Steps	{0}	0
Batch Size	{16, 8}	16

Table D2: Hyper-parameter search space and selected values. We performed a grid-search on a total of 48 different settings. The best choices per hyper-parameter can be seen in the last column.

DA with instances from BIOMRC	
ID	Instance
16061304	**BIOMRC question:** Prognosis of 6644 resected [MASK] in Japan: a Japanese lung cancer registry study. **BIOMRC snippet:** Otherwise, the present TNM staging system seemed to well characterize the stage-specific prognosis in non-small cell lung cancer . **BIOMRC answer:** non-small cell lung cancer
19823942	**BIOMRC question:** Systolic versus diastolic cardiac function variables during [MASK] treatment for breast cancer . **BIOMRC snippet:** epirubicin induces considerable decrease in left ventricular ejection fraction and a high risk of CHF. **BIOMRC answer:** epirubicin
22457372	**BIOMRC question:** Pre-operative education and counselling are associated with [MASK] following carotid endarterectomy: a randomized and open-label study. **BIOMRC snippet:** AIM: To investigate the effect of pre-operative visits and counselling by intensive care unit (intensive care unit) nurses on Patients 's anxiety symptoms following carotid endarterectomy. **BIOMRC answer:** anxiety symptoms

Table D3: Training instances extracted from BIOMRC. Each instance is a triple containing a cloze-style question, a snippet, and the span of the snippet answering the question.

DA via question back-translation	
ID	Instance
8699317	**Pivot language:** French **BIOASQ question:** Which is the gene mutated in type 1 neurofibromatosis? **Back-translated Question:** What is the mutated gene in type 1 neurofibromatosis? **BIOASQ snippet:** An NF1 gene was identified as a gene whose loss of function causes an onset of human disorder, neurofibromatosis type I. **BIOASQ answer:** NF1
11816795	**Pivot language:** Spansih **BIOASQ question:** Which is the primary protein component of Lewy bodies? **Back-translated question:** What is the main protein component of Lewy bodies? **BIOASQ snippet:** The protein alpha-synuclein appears to be an important structural component of Lewy bodies, an observation spurred by the discovery of point mutations in the alpha-synuclein gene linked to rare cases of autosomal dominant PD. **BIOASQ answer:** alpha-synuclein
3056562	**Pivot language:** German **BIOASQ question:** Which type of urinary incontinence is diagnosed with the Q tip test? **Back-translated question:** What type of urinary incontinence does the Q tip test diagnose? **BIOASQ snippet:** Simple clinical tests for support of the urethrovesical junction, such as the Q tip test, are non-specific in patients with stress urinary incontinence. **BIOASQ answer:** stress urinary incontinence

Table D4: Training instances generated via back-translation of BIOASQ questions using French, Spanish, and German as a pivot language. A generated instance contains a back-translated question and the corresponding BIOASQ snippet and answer.

DA via snippet back-translation	
ID	Instance
8699317	**Pivot language:** French **BIOASQ question:** Which is the protein implicated in Spinocerebellar ataxia type 3? **BIOASQ snippet:** Ataxin-3 (AT3) is the protein that triggers the inherited neurodegenerative disorder spinocerebellar ataxia type 3 when its polyglutamine (polyQ) stretch close to the C-terminus exceeds a critical length **Back-translated snippet:** Ataxin-3 (AT3) is the protein that triggers spinocerebellar ataxia type 3 in inherited neurodegenerative disorder when its polyglutamine (polyQ) stretches near the C-terminus exceeds a critical length. **BIOASQ answer:** Ataxin-3
16232326	**Pivot language:** Spanish **BIOASQ question:** Which gene is responsible for the development of Sotos syndrome? **BIOASQ snippet:** Haploinsufficiency of the NSD1 gene has been implicated as the major cause of Sotos syndrome, with a predominance of microdeletions reported in Japanese patients **Back-translated snippet:** NSD1 gene haploinsufficiency has been implicated as the main cause of Sotos syndrome, with a predominance of microdeletions reported in Japanese patients. **BIOASQ answer:** NSD1 gene
11154546	**Pivot language:** German **BIOASQ question:** Abnormality in which vertebral region is important in the Bertolotti's syndrome? **BIOASQ snippet:** Repeated fluoroscopically guided injections implicated a symptomatic L6-S1 facet joint contralateral to an anomalous lumbosacral articulation. **Back-translated snippet:** Repeated fluoroscopic injections implied a symptomatic L6-S1 facet joint contralateral to an abnormal lumbosacral articulation. **BIOASQ answer:** lumbosacral

Table D5: Training instances generated via back-translation of BIOASQ snippets using French, Spanish, and German as a pivot language. A generated instance contains a back-translated snippet and the corresponding BIOASQ question and answer.

DA via Information Retrieval	
ID	Instance
25941473	**BIOASQ question:** Which is the neurodevelopmental disorder associated to mutations in the X- linked gene mecp2? **Retrieved snippet:** Genotype-specific effects of Mecp2 loss-of-function on morphology of Layer V pyramidal neurons in heterozygous female Rett syndrome model mice. **BIOASQ answer:** rett syndrome
28708333	**BIOASQ question:** Which is the molecular target of the immunosuppressant drug Rapamycin? **Retrieved snippet** Conversion from calcineurin inhibitors to mTOR inhibitors as primary immunosuppressive drugs in pediatric heart transplantation. **BIOASQ answer:** mtor
29767248	**BIOASQ question:** What is the target of the drug Olaparib? **Retrieved snippet:** Mechanistically, dual blockade of PI3K and PARP in ARID1A-depleted gastric cancer cells significantly increased apoptosis detected by flow cytometry, and induced DNA damage by immunofluorescent staining. **BIOASQ answer:** parp

Table D6: Training instances generated via IR. A BIOASQ question is used as the query to retrieve PUBMED documents. For each snippet of the retrieved documents that contains the answer, we generate a new training triplet consisting of the BIOASQ question, the snippet and the BIOASQ answer.

DA with word substitution based on WORD2VEC	
ID	Instance
27965160	**BIOASQ question:** Sclerostin regulates what process? **BIOASQ snippet:** Sclerostin is a soluble antagonist of Wnt/b-catenin signaling secreted primarily by osteocytes. Current evidence indicates that sclerostin likely functions as a local/paracrine regulator of bone metabolism rather than as an endocrine hormone. **Snippet after WORD2VEC substitution:** sclerostin is a soluble agonist of wnt-b catenin signaling secreted mainly by osteocytes current evidence suggests that sclerostin likely functions as a localparacrine regulator of bone metabolism rather than as an endocrine hormone **BIOASQ answer:** bone metabolism
22003227	**BIOASQ question:** Which microRNA is the mediator of the obesity phenotype of patients carrying 1p21.3 microdeletions? **BIOASQ snippet:** The study also demonstrated significant enrichment of miR-137 at the synapses of cortical and hippocampal neurons, suggesting a role of miR-137 in regulating local synaptic protein synthesis machinery. CONCLUSIONS: This study showed that dosage effects of MIR137 are associated with 1p21.3 microdeletions and may therefore contribute to the ID phenotype in patients with deletions harbouring this miRNA. **Snippet after WORD2VEC substitution:** the study also demonstrated significant enrichment of mir 137 at the synapses of cortical and hippocampal neurons indicating a implication of mir 137 in regulating local synaptic protein synthesis machinerybr-bconclusionsb this study showed that dosage effects of mir137 are associated with 2q223 microdeletions and might hence contribute to the id phenotype in patients with microinsertions harbouring this micro-rna **BIOASQ answer:** MIR137
21546092	**BIOASQ snippet:** Beck's Medical Lethality Scale (BMLS) was administered to assess the degree of medical injury, and the SAD PERSONS mnemonic scale was used to evaluate suicide risk. **BIOASQ question:** What is evaluated with the SAD PERSONS scale? **Snippet after WORD2VEC substitution:** becks medical lethality scale bmls was administered to evaluate the degree of medical injury and the sad people domain-general scale was utilized to investigate suicide risk **BIOASQ answer:** suicide risk

Table D7: Training instances generated via word substitution based on WORD2VEC. We randomly select at most 10 words of a BIOASQ snippet and substitute each word w_i with its most similar word w_j from the vocabulary of the WORD2VEC model. Highlights of the same color indicate substituted words and the corresponding substitutions.

DA with word substitution based on BIOLM	
ID	Instance
22140526	**BIOASQ question:** Which gene is responsible for red hair? **BIOASQ snippet:** The association signals at the MC1R gene locus from CDH were uniformly more significant than traditional GWA analyses. The CDH test will contribute towards finding rare LOF variants in GWAS and sequencing studies. **BIOASQ snippet after BIOLM substitution:** The association signals at the MC1R 1 identified from CDH were significantly more significant than traditional association analyses. The proposed findings will contribute towards detecting novel risk variants in GWAS and sequencing studies. **BIOASQ answer:** MC1R
26917818	**BIOASQ question:** Dinutuximab is used for treatment of which disease? **BIOASQ snippet:** CONCLUSIONS Dinutuximab is the first anti-GD2 monoclonal antibody approved in combination with GM-CSF, IL-2, and RA for maintenance treatment of pediatric patients with high-risk neuroblastoma who achieve at least a partial response to first-line multiagent, multimodality therapy. **BIOASQ snippet after BIOLM substitution:** CONCLUSIONS Dinutuximab is the first human monoclonal antibody approved in combination with recombinant IL-2, and dexamethasone for maintenance treatment of pediatric patients with high-risk neuroblastoma who achieve at least a partial response to prior multiagent, standard therapy. **BIOASQ answer:** neuroblastoma
27789693	**BIOASQ question:** Which database associates human noncoding SNPs with their three-dimensional interacting genes? **BIOASQ snippet:** 3DSNP: a database for linking human noncoding SNPs to their three-dimensional interacting genes. **BIOASQ snippet after BIOLM substitution:** 3DSNP: a method for linking functional GWAS SNPs to their three-dimensional structural structures **BIOASQ answer:** 3DSNP

Table D8: Training instances generated via word substitution based on BIOLM.We randomly select at most 10 words of a BIOASQ snippet and we substitute each word w_i with the most probable word w_j suggested by BIOLM after masking w_i. Highlights of the same color indicate substituted words and the corresponding substitutions.

DA via Question Generation using BIOASQ snippets	
ID	Instance
21159650	**Generated question:** What enzyme inhibits cullin-RING E3 ubiquitin ligases? **BIOASQ snippet:** MLN4924 is a first-in-class experimental cancer drug that inhibits the NEDD8-activating enzyme, thereby inhibiting cullin-RING E3 ubiquitin ligases and stabilizing many cullin substrates **Generated answer:** NEDD8
17333537	**Generated question:** What type of RNA triggers silencing of inactivation in eutherian mammals? **BIOASQ snippet:** In eutherian mammals X inactivation is regulated by the X-inactive specific transcript (Xist), a cis-acting non-coding RNA that triggers silencing of the chromosome from which it is transcribed **Generated answer:** chromosome
16800744	**Generated question:** What is the human tissue kallikrein family of? **BIOASQ snippet:** The human tissue kallikrein family of serine proteases (hK1-hK15 encoded by the genes KLK1-KLK15) is involved in several cancer-related processes. **Generated answer:** serine proteases

Table D9: Training instances generated using T5. Given a BIOASQ snippet T5selects a span of the snippet and generates a question that can be answered by that span. We select spans different than the ones used in BIOASQ.

DA via Question Generation using random snippets from random PUBMED abstracts	
ID	Instance
26935709	**Generated question:** What can be isolated or in combination with accompanying deformities occurring in the forefoot and/or hindfoot? **PUBMED snippet:** Symptoms can be isolated or in combination with accompanying deformities occurring in the forefoot and/or hindfoot. **Generated answer:** Symptoms
29260288	**Generated question:** What supplementation has been integrated into our practice? **PUBMED snippet:** Vitamin D supplementation has been integrated into our current practice. **Generated answer:** Vitamin D
30706485	**Generated question:** What were connected to a volume-cycled ventilator after sedation, analgesia and endotracheal intubation? **PUBMED snippet:** After sedation, analgesia and endotracheal intubation, pigs were connected to a volume-cycled ventilator. **Generated answer:** pigs

Table D10: Training instances generated using T5. Given a random snippet from a random PUBMED article T5selects a span of the snippet and generates a question that can be answered by that span.

DA by adding context	
ID	Instance
15149039	**BIOASQ question:** Which metabolite activates AtxA? **BIOASQ snippet:** Transcription of the major Bacillus anthracis virulence genes is triggered by CO2, a signal mimicking the host environment. **BIOASQ snippet with additional context:** Transcription of the major Bacillus anthracis virulence genes is triggered by CO2, a signal mimicking the host environment. A 182-kb plasmid, pXO1, carries the anthrax toxin genes and the genes responsible for their regulation of transcription, namely atxA and, pagR, the second gene of the pag operon. AtxA has major effects on the physiology of B. anthracis. It coordinates the transcription activation of the toxin genes with that of the capsule biosynthetic enzyme operon, located on the second virulence plasmid, pXO2. In rich medium, B. anthracis synthesises alternatively two S-layer proteins (Sap and EA1). **Answer:** CO2
16757427	**BIOASQ question:** What tyrosine kinase, involved in a Philadelphia- chromosome positive chronic myelogenous leukemia, is the target of Imatinib (Gleevec)? **BIOASQ snippet:** Imatinib was developed as the first molecularly targeted therapy to specifically inhibit the BCR-ABL kinase in Philadelphia chromosome (Ph)-positive chronic myeloid leukemia (CML). **BIOASQ snippet with additional context:** The second generation of BCR-ABL tyrosine kinase inhibitors. Imatinib was developed as the first molecularly targeted therapy to specifically inhibit the BCR-ABL kinase in Philadelphia chromosome (Ph)-positive chronic myeloid leukemia (CML). Because of the excellent hematologic and cytogenetic responses, imatinib has moved toward first-line treatment for newly diagnosed CML. However, the emergence of resistance to imatinib remains a major problem in the treatment of Ph-positive leukemia. Several mechanisms of imatinib resistance have been identified, including BCR-ABL gene amplification that leads to overexpression of the BCR-ABL protein, point mutations in the BCR-ABL kinase domain that interfere with imatinib binding, and point mutations outside of the kinase domain that allosterically inhibit imatinib binding to BCR-ABL. **Answer:** BCR-ABL

Table D11: Training instances generated by adding context around the original BIOASQ snippet. In the generated snippet the original one is highlighted.