Dense Passage Retrieval with Backtranslation-based data Augmentation for ODQA

Seongjun, Yang
KAIST
seongjunyang@kaist.ac.kr

Seungwoo, Ryu
KAIST
swryu94@kaist.ac.kr

Abstract

Open-domain question answering (ODQA) concentrates on finding good passage retrieval algorithms to find candidate documents, and sparse retrieval models such as TF-IDF and BM25 are generally used. In this paper, we show that our dense retrieval model whose questions are augmented with back-translation outperforms a baseline dense retrieval model, DPR. When evaluated on one of the ODQA datasets, Natural Question (NQ), our back-translation-augmented model (**DPR-BA**) with low temperature (T=0.1) outperforms DPR about 2% for both retrieval and reader outputs in Top-100 accuracy, and generated questions are highly qualitative. In addition, even syntactically or semantically deformed questions from back-translation actually improve the performance of QA prediction.

1 Instructions

Open-domain question answering (ODQA) is a task to find a relevant answer on a given question in a large amount of documents without being provided a pre-defined subject. Although early ODQA systems were devised in a complex manner (Moldovan et al., 2003), modern ODQA methodologies suggest a simple two-stream architectures: *retrieval* and *reader*. A retrieval finds candidate documents which are probable to contain an answer to a question, and a reader extracts an answer corresponding to a given question from extracted documents.

Retrieval in ODQA usually uses word-based sparse retrieval methods such as TF-IDF and BM25 (Robertson and Zaragoza, 2009). It counts the appearance of words with inverted indices and fills per-document vectors with statistic values obtained from each method. For example, TF-IDF calculates Term-Frequency (TF) and Inverse Document Frequency (IDF) to obtain appropriate weights to be multiplied on inverted words. On the contrary, dense retrieval methods fill the vectors in a dense way using a Neural Network-based approach. This approach has two advantages over sparse retrieval. First, it is good at capturing syntactic and semantic meaning from documents. For example, if a question is suggested as '*What did Albert Einstein win the Nobel Prize for?*', a dense vector recognizes that it is syntactically wrong to answer as '*Law the effect of photoelectic the.*', instead of '*The law of the photoelectric effect.*'. In addition, dense vector might perceive that '*The principle of the photoelectric effect.*' has a similar meaning with the original answer. Second, dense vectors are learnable. Unlike sparse retrieval, vector is not fixed per document. In an end-to-end manner, encoder optimizes its elements of vectors in such a way that a vector expresses its document in a best.

It is generally believed that sparse retrieval performs better than dense retrieval. However, ORQA (Lee et al., 2019) outperform BM25 by using *de facto* methodology of these days, *Pretraining*. ORQA suggested additional pretraining task, Inverse Cloze Task (ICT), which takes a selected sentence from a document as a query, and predicts its context. Further, REALM (Guu et al., 2020) augments pretrained language model by introducing a huge-scale additional encoder, Neural Knowledge Retrieval, which *explicitly* encourages helpful retrieval and penalizes unhelpful retrieval. However, such additional pretraining steps require huge computation. On the contrary, Dense Passage Retriever (DPR) (Karpukhin et al., 2020) eliminated additional pretraining step by introducing simple dual Bert-base-uncased encoders on question and passage.

In this paper, we deal with this question: is it possible to improve performance with data augmentation without introducing an additional computationally heavy pretraining task? We leverage the same dual BERT-encoders architectures from DPR, but augmented the question set twofold by using a *back-translation* methodology. Through

meticulous back-translation experiments, our conclusion is simple: Highly qualitative questions augmented from back-translation helps improve the performance of ODQA task in both retrieval and reader stage. Our DPR-BA model with T=0.1 always outperforms DPR, for example, not only on *Dense Retrieval* accuracy (67.1 % vs. 66 % in Top-100 accuracy), but also on *Reader* EM (33.2 % vs. 29.8 % in Top-100 EM) in NQ dataset.

Our model has two key contributions on ODQA task. First, we re-identify that training with more semantically similar data increases model's performance like the performance of the language model improves as data size increases (Kaplan et al., 2020). Although the questions generated from back-translation process are synthetic, if questions can imitate the syntactic and semantic meaning of the real questions, they help the model to retrieve relevant documents. Performance improvement on document retrieval is passed to the performance improvement of reader. Second, even semantically or syntactically deformed questions generated from back-translation are also helpful in improving overall reader performance. Regardless of the size of temperature, our model improves EM score.

2 Background

Back-translation Neural Machine Translation (NMT) is an trial to train a large Neural Network based model which can understand the meaning of original sentence and then translate it to a sentence in a target language properly (Bahdanau et al., 2014). Back-translation was first introduced to improve the frequency of a sentence generated from NMT model with monolingual data (Sennrich et al., 2016). It is a process of generating *synthetic source sentence* refering to the monolingual target sentence. We focus on using transformer-based back-translation model which samples a synthetic source through sampling from model distribution and adding some noise to an output obtained from beam search (Edunov et al., 2018).

Open-domain question answering (ODQA) ODQA is a task of selecting relevant documents (*retreival*) and then giving a correct answer (*reader*) without any specifically-provided document regarding a question. For example, if the question is given as "`Where is the capital city of South Korea?`", a well-trained ODQA model should give an correct answer "`Seoul`" in the pool of many capital city-unrelated documents. The overall framework for ODQA is illustrated on (FIGURE 1). We focus on *Dense Retrieval* method which can capture syntactic and semantic meaning of the words in a document. Dense Passage Retrieval uses the the same dual BERT-encoders architecture with *In-batch negatives* setting in *Retrieval* stage.

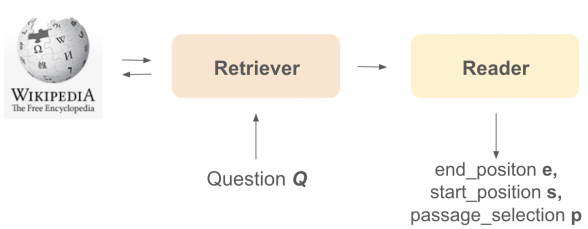

Figure 1: Open domain question answering

3 Approach

We posit that Retriever with auxiliary question generation can capture more semantically relevant documents. So, we expect more higher accuracy in end-to-end QA. By the way, it is important to generate semantically similar but syntactically diverse questions. So, we use back-translation for question generation. generated questions from back-translation provide targeted inductive bias to the model while keeping the meaning of source sentences (Gunel et al., 2021).

In our task, we use back-translation model trained with WMT-19 En-De (Ott et al., 2019). By the way, back-translation creates more noisy samples when we increase temperature hyperparameter in random sampling (Gunel et al., 2021). it means semantically incorrect sentences can be generated in high temperature. So, we generate questions with various temperatures. Then, in the Dense Retriever model, generated questions Q' are inserted when calculating similarity of question vector Q and passage vector p.

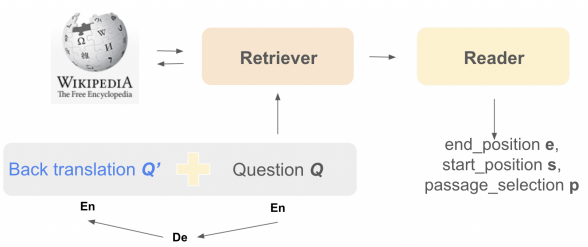

Figure 2: New approach

4 Experiments

Hyperparameter Detailed hyperparameter setting of baselines and DPR-BA is stated at TABLE 1. In our task, we use 2 questions and train models with 4 temperature settings because of resource limits. We also followed the hyperparameter setting suggested by Github of professor Minjoon Seo.

Dense retriever setting	
Pretrained_model	bert-base-uncased
encoder_model_type	hf_bert
max_grad_norm	2.0
seed	12345
sequence_length	256
warmup_steps	1237
batch_size	12
do_lower_case	Y
learning_rate	2e-05
num_train_epochs	15
dev_batch_size	12
val_av_rank_start_epoch	10
Task Coefficient	
number_of_qustion	2
Temperature (T)	{0.1, 0.3, 0.5, 0.7}

Table 1: Hyperparameter settings

Baseline We compare DPR-BA on two baseline settings. One is the model provided by TA (BASELINE). However, our utility of GPU resources is limited, so we run the DPR-BA with 15 train epochs and 10 var_av_rank_start_epochs. Then, We rerun the baseline code under the diminished hyper-parameter setting likewise, as stated in (TABLE 1), We call this model as (NEW_BASE).

Evaluation We measured the performance of our model in a same measurement setting with DPR. We choose *Top-k* accuracy with various k values as a metric which measures the performance of passage retrieval and Exact Matching (EM) in Question Answering. We control temperature parameter T in back-translation to adjust the diversity of generated questions. So, we also represent some qualitative examples generated from back-translation process to observe how semantically similar sentences are generated.

Qualitative examples of generated questions (TABLE 2) shows the comparison between the real question and the questions generated from back-translation. In case of the question in Original 1, all the generated questions maintain the semantics of original question regardless of temperature parameter, but somewhat syntactically deformed counterparts are generated in a consistent way. In case of the question in Original 2, syntactic and semantic meanings are well preserved when T is small. However, as temperature parameter increases, a generated sample loses its semantics and syntactics gradually. When T is 0.7, the generated sample is completely different from the original question in both syntactic and semantic perspective. Further, in case of the question in Original 3, semantic and syntactic are totally deformed even at the early stage of temperature parameter. Although T is really small, back-translated process has some difficulty in understanding *showshank redemption* itself as a independent movie name. We hypothesize back-translation itself suffers some difficulty in language understanding when it faces complex sentences.

Original_1	what's the legal marriage age in new york
T=0.1	What is the legal age of marriage in New York?
T=0.3	What is the legal age of marriage in New York?
T=0.5	What is the legal age of marriage in New York?
T=0.7	What is the legal age of marriage in New York?
Original_2	what is the channel number for cartoon network on spectrum
T=0.1	What is the channel number for cartoon network on the spectrum
T=0.3	What is the channel number for cartoon network on the spectrum
T=0.5	What is the channel number for the cartoon network on the frequency spectrum
T=0.7	What is the number of the radio station for the cartoon network on the frequency spectrum?
Original_3	who sings the marriage of figaro in shawshank redemption
T=0.1	who sings the wedding of Figaro in a shroud
T=0.3	who sings the wedding of Figaro in a shroud
T=0.5	who sings the wedding of Figaro in a bowl apron
T=0.7	singing the marriage of Figaro in an apron

Table 2: Qualitative example of generated questions

Passage retrieval accuracy (TABLE 3) shows passage retrieval accuracy. Our back-translation based augmentation model consistently shows the higher accuracy than DPR(NEW_BASE) when T is 0.1. However, as shown in TABLE 2, there are many cases when back-translation itself suffers difficulty to capture meaning of question, as temperature parameter increases, accuracy is decreased. We hypothesize generating questions in a similar distribution encourages document-retrieval in relevant.

	Baseline	New_base	T=0.1	T=0.3	T=0.5	T=0.7
Top-1	**35.39**	31.67	32.65	30.26	28.37	29.01
Top-20	**58.62**	54.79	56.04	53.06	52.11	53.78
Top-40	**63.38**	60.01	61.23	58.57	57.70	59.44
Top-60	**65.83**	62.83	63.98	61.50	60.72	62.37
Top-80	**67.47**	64.63	65.77	63.47	62.74	64.28
Top-100	**68.68**	65.99	67.06	64.92	64.20	65.68

Table 3: Passage retrieval accuracy

End to end QA result (TABLE 4) shows an end to end QA result in reader function with respect to Exact Matching(EM) score. When generated question through back-translation is similar with the original question (T=0.1), QA performance of DPR-BA always outperforms NEW_BASE. However, even when generated questions are deformed semantically and syntactically by high temperature, they are also helpful at predicting correct answer in QA. We hypothesize that they are helpful although they are deformed and semantically unrelated questions.

	Baseline	New_base	T=0.1	T=0.3	T=0.5	T=0.7
Top-10	29.50	29.00	**31.29**	30.16	29.46	30.01
Top-20	30.33	29.06	**32.60**	31.45	31.16	31.62
Top-40	29.78	29.78	**33.15**	32.44	31.92	32.73
Top-50	29.78	29.78	**33.15**	32.44	31.92	32.73
Top-80	29.78	29.78	**33.15**	32.44	31.92	32.73
Top-100	29.78	29.78	**33.15**	32.44	31.92	32.73

Table 4: End to end QA result

5 Conclusion

In this work, we augmented the question set for retrieval twofold using back-translation method. Generated questions are helpful at dense retrieval only when they are semantically and syntactically similar with the original question, but in end-to-end manner, deformed questions are always helpful at QA prediction in the end. However, DPR-BA is memory and computationally-heavy. DPR takes a day for training using only 13000MiB, on the other hand, our model takes $2 \sim 3$ days using about 19000MiB memory usage. So, we were not able to train DPR-BA with many hyperparameter settings due to resource constraints. For future works, we need to train DPR-BA with various settings. In addition, we expect a more memory-efficient question augmentation method is applied in DPR.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Dan Moldovan, Chris Clark, Sanda Harabagiu, and Steven J Maiorano. 2003. Cogex: A logic prover for question answering. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 166–172.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

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. Association for Computational Linguistics.