Context-Aware Answer Extraction in Question Answering

Yeon Seonwoo[†], **Ji-Hoon Kim**[‡§], **Jung-Woo Ha**[‡§], **Alice Oh**[†]
[†]KAIST
[‡]NAVER AI LAB, [§]NAVER CLOVA
yeon.seonwoo@kaist.ac.kr
{genesis.kim,jungwoo.ha}@navercorp.com
alice.oh@kaist.edu

Abstract

Extractive QA models have shown very promising performance in predicting the correct answer to a question for a given passage. However, they sometimes result in predicting the correct answer text but in a context irrelevant to the given question. This discrepancy becomes especially important as the number of occurrences of the answer text in a passage increases. To resolve this issue, we propose **BLANC** (**BL**ock **A**ttentio**N** for **C**ontext prediction) based on two main ideas: context prediction as an auxiliary task in multi-task learning manner, and a block attention method that learns the context prediction task. With experiments on reading comprehension, we show that BLANC outperforms the state-of-the-art QA models, and the performance gap increases as the number of answer text occurrences increases. We also conduct an experiment of training the models using SQuAD and predicting the supporting facts on HotpotQA and show that BLANC outperforms all baseline models in this zero-shot setting.

1 Introduction

Question answering tasks require a high level of reading comprehension ability, which in turn requires a high level of general language understanding. This is why the question answering (QA) tasks are often used to evaluate language models designed to be used in various language understanding tasks. Recent advances in contextual language models brought on by attention (Hermann et al., 2015; Chen et al., 2016; Seo et al., 2017; Tay et al., 2018) and transformers (Vaswani et al., 2017) have led to significant improvements in QA, and these improvements show that better modeling of contextual meanings of words plays a key role in QA.

While these models are designed to select answer-spans in the relevant contexts from given passages, they sometimes result in predicting the correct answer text but in contexts that are irrelevant to the given questions. Figure 1 shows an example passage where the correct answer text appears multiple times. In this example, the only answer-span in the context relevant to the given question is the first occurrence of the "prefrontal cortex" (in blue), and all remaining occurrences of the answer text (in red) show incorrect predictions. Figure 2 shows quantitatively, the discrepancy between predicting the correct answer text versus predicting the correct answer-span. Using BERT (Devlin et al., 2019) trained on curated NaturalQuestions (Fisch et al., 2019), we show the results of extractive QA task using exact match (EM) and Span-EM. EM only looks for the text to match the ground truth answer, whereas Span-EM additionally requires the span to be the same as the ground truth answer-span. Figure 2 shows that BERT finds the correct answer text more than it finds the correct answer-spans, and this proportion

Passage: Some of the most developmentally significant changes in the brain occur in the **prefrontal cortex**, which is involved in decision making and cognitive control, as well as other higher cognitive functions. ... pruning in the **prefrontal cortex** increases, improving the efficiency of information processing, and neural connections between the **prefrontal cortex** and other regions of the brain are strengthened. ... Specifically, developments in the dorsolateral **prefrontal cortex** are important for controlling impulses and planning ahead, while development in the ventromedial **prefrontal cortex** is important for decision making.

Question: Which part of the brain is involved in decision making and cognitive control?

Answer: prefrontal cortex

Figure 1: Example passage, question, and answer triple. This passage has multiple spans that are matched with the answer text. The first occurrence of "prefrontal cortex" is the only answer-span within the context of the question.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2418–2428,
November 16–20, 2020. ©2020 Association for Computational Linguistics

of wrong predictions increases as the number of occurrences of answer text in a passage increases.

Tackling this problem is very important in more realistic datasets such as NaturalQuestions (Kwiatkowski et al., 2019), where the majority of questions have more than one occurrence of the answer text in the passage. This is in contrast with the SQuAD dataset, where most questions have a single occurrence of the answer. These details of the SQuAD (Rajpurkar et al., 2016), NewsQA, and NaturalQuestions datasets (Fisch et al., 2019) are shown in Figure 3.

To address this issue, we define context prediction as an auxiliary task and propose a block attention method, which we call **BLANC** (**BL**ock **A**ttentio**N** for **C**ontext prediction) that explicitly forces the QA model to predict the context. We design the context prediction task to predict soft-labels which are generated from given answer-spans. The block attention method effectively calculates the probability of each word in a passage being included in the context with negligible extra parameters and inference time. We provide the implementation of BLANC publicly available [1].

Adding context prediction and block attention enhances BLANC to correctly identify context related to a given question. We conduct two types of experiments to verify the context differentiation performance of BLANC: extractive QA task, and zero-shot supporting facts prediction. In the extractive QA task, we show that BLANC significantly increases the overall reading comprehension performance, and we verify the performance gain increases as the number of answer texts in a passage increases. We verify BLANC's context-aware performance in terms of generalizability in the zero-shot supporting facts prediction task. We train BLANC and baseline models on SQuAD1.1. and perform zero-shot supporting facts (supporting sentences in passages) prediction experiment on HotpotQA dataset (Yang et al., 2018). The results show that the context prediction performance that the model has learned from one dataset is generalizable to predicting the context of an answer to a question in another dataset.

Contributions in this paper are as follows:

- We show the importance of correctly identifying the answer-span to improving the model performance on extractive QA.

[1] https://github.com/yeonsw/BLANC

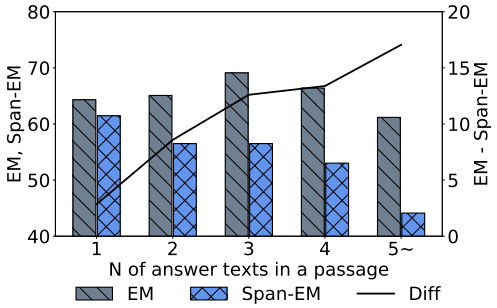


Figure 2: EM (text-matching) and Span-EM (span matching) of BERT on the groups divided by the number of answer text occurrences in a passage. Note: The difference for $N = 1$ results from post-processing steps (removing articles) in EM evaluation.

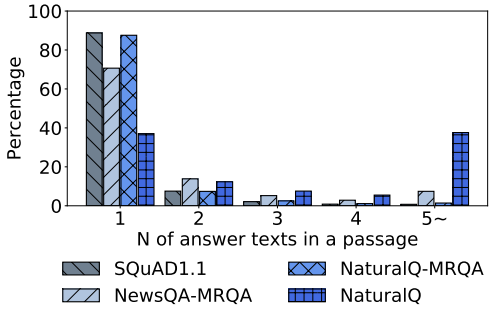


Figure 3: Proportions of questions with various numbers of the answer text in a passage. SQuAD has only a few examples for ($n \geq 5$), while NaturalQuestions has a large proportion.

- We show that context prediction task plays a key role in the QA domain.

- We propose a new model BLANC that resolves the discrepancy between answer text prediction and answer-span prediction.

2 Related Work

Evidence in the form of documents, paragraphs, and sentences, has been shown to be necessary and effective in predicting the answers in open-domain QA (Chen et al., 2017; Wang et al., 2018; Das et al., 2018; Lee et al., 2019) and multi-hop QA (Yang et al., 2018; Min et al., 2019b; Asai et al., 2020). One problem of identifying evidence in answering questions is the expensive cost in labeling the evidence. Self-labeling with simple heuristics can be a solution to this problem, as shown in Choi et al. (2017); Li et al. (2018). Self-training is another solution, as presented in Niu et al. (2020). In this paper, we propose self-generating soft-labeling method to indicate support words of answer texts, and train BLANC with the soft-labels.

Related but different from our work, Swayamdipta et al. (2018) and Min et al. (2019a) predict the answer-span when only the answer texts are provided and the ground truth answer-spans are not. Swayamdipta et al. (2018) designs a model that benefits from aggregating information from multiple mentions of the answer text in predicting the final answer. Min et al. (2019a) approach the problem of the lack of ground truth answer-spans with latent modeling of candidate spans. Both of these papers tackle the problem of identifying the correct answer among multiple mentions of the answer text in datasets without annotations of the correct answer-spans. Our work solves a different problem from the above-mentioned papers in that the golden answer-spans are provided.

3 Model

We propose BLANC based on two novel ideas: soft-labeling method for the context prediction and a block attention method that predicts the soft-labels. Two important functionalities of BLANC are 1) calculating the probability that a word in a passage belongs to the context, which is in latent, and 2) enabling the probability to reflect spatial locality between adjacent words. We provide an overall illustration of BLANC in Figure 4.

3.1 Notations

In this section, we define the notations and the terms used in our study. We denote a word at index i in a passage with w_i. We define the context of a given question as a segment of words in a passage and denote with $\mathcal{C}$. In our circumstance, the context is latent. We denote the start and end indices of a context with s_c and e_c. Training a block attention model to predict the context requires the labeling process for the latent context, and we define two probabilities for that, $p_{\text{soft}}(\mathrm{w}_i \in \mathcal{C})$ and $p(\mathrm{w}_i \in \mathcal{C})$. $p_{\text{soft}}(\mathrm{w}_i \in \mathcal{C})$ represents the self-generated soft-label that we assume as ground truth of the context, and $p(\mathrm{w}_i \in \mathcal{C})$ is a block attention model's prediction. We denote the start and end indices of a labeled answer-span with s_a and e_a.

3.2 Soft-labeling for latent context $\mathcal{C}$

We assume words near an answer-span are likely to be included in the context of a given question. From our assumption, we define the probability of words belong to the context, $p_{\text{soft}}(\mathrm{w}_i \in \mathcal{C})$, which is used as a soft-label for the auxiliary context prediction task. To achieve this, we hypothesize the words in an answer-span are included in the context and make the probability of adjacent words decrease with a specific ratio as the distance between answer-span and a word increases. The soft-label for the latent context is as follows:

$$p_{\text{soft}}(\mathrm{w}_i \in \mathcal{C}) = \begin{cases} 1.0 & \text{if } i \in [s_a, e_a] \\ q^{|i-s_a|} & \text{if } i < s_a \\ q^{|i-e_a|} & \text{if } i > e_a, \end{cases} \quad (1)$$

where $0 \leq q \leq 1$, and q is a hyper-parameter for the decreasing ratio as the distance from a given answer-span. For computational efficiency, we apply (1) to words bounded by certain window-size only, which is a hyper-parameter, on both sides of an answer-span. This results in assigning $p_{\text{soft}}(\mathrm{w}_i \in \mathcal{C})$ to 0 for the words outside the segment bounded by the window-size.

3.3 Block Attention

Block attention model calculates $p(\mathrm{w}_i \in \mathcal{C})$ to predict the soft-label, $p_{\text{soft}}(\mathrm{w}_i \in \mathcal{C})$, and localizes the correct index of an answer-span with $p(\mathrm{w}_i \in \mathcal{C})$. We embed spatial locality of $p(\mathrm{w}_i \in \mathcal{C})$ to block attention model with the following steps: 1) predicting the start and end indices of context, $p(i = s_c)$ and $p(i = e_c)$, and 2) calculating $p(\mathrm{w}_i \in \mathcal{C})$ with cumulative distribution of $p(i = s_c)$ and $p(i = e_c)$. In the first step, at predicting the start and end indices, all encoder models that produce vector representation of words in a passage are compatible with the block attention model. In this paper, we apply the same structure of the answer-span classification layer used in the transformer model (Devlin et al., 2019) to our context words prediction layer.

$$\boldsymbol{H} = \text{Encoder}(\text{Passage}, \text{Question}) \quad (2)$$

Here, we denote $\boldsymbol{H}$ as output vectors of transformer encoder and $\boldsymbol{H}_j$ as output vector of w_j. From $\boldsymbol{H}$, we predict the start and end indices of the context:

$$\begin{aligned} p(i = s_c) &= \frac{\exp(\boldsymbol{W_c} \boldsymbol{H}_i + b_s^c)}{\sum_j \exp(\boldsymbol{W_c} \boldsymbol{H}_j + b_s^c)}, \\ p(i = e_c) &= \frac{\exp(\boldsymbol{V_c} \boldsymbol{H}_i + b_e^c)}{\sum_j \exp(\boldsymbol{V_c} \boldsymbol{H}_j + b_e^c)}, \end{aligned} \quad (3)$$

where $\boldsymbol{W_c}$, $\boldsymbol{V_c}$, b_s^c, and b_e^c represent weight and bias parameters for context prediction layer. We calculate $p(\mathrm{w}_i \in \mathcal{C})$ as multiplication of the probability

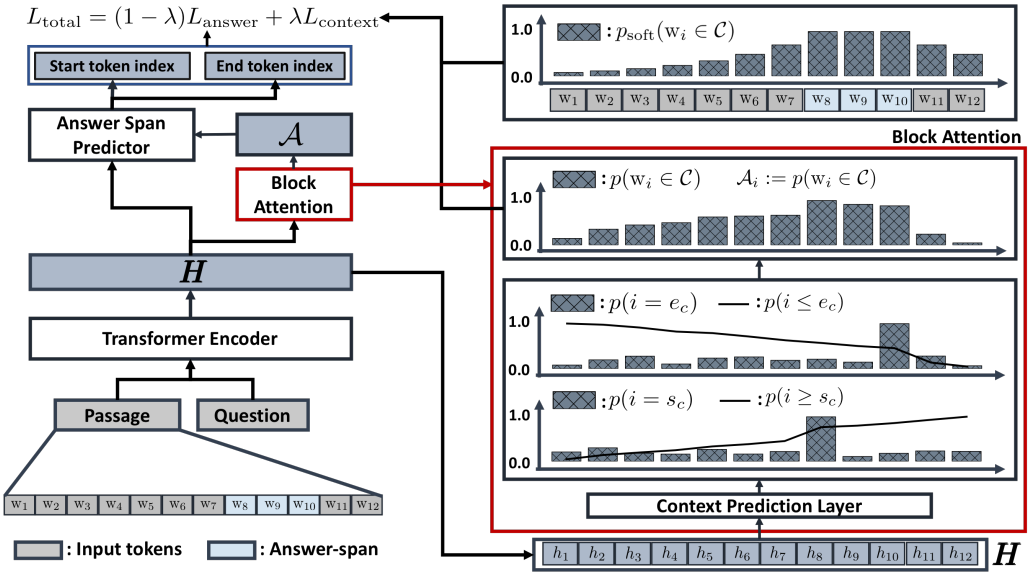


Figure 4: Schematic visualization of BLANC. Block attention model takes contextual vector representations from transformer encoder and predicts context words of an answer, $p(\text{w}_i \in \mathcal{C})$. We define loss function for context words with the prediction, $p(\text{w}_i \in \mathcal{C})$ and the self-generated soft-label $p_{\text{soft}}(\text{w}_i \in \mathcal{C})$ defined in (1). Answer-span predictor takes $p(\text{w}_i \in \mathcal{C}$ and $\boldsymbol{H}$ to predict an answer-span. We optimize our model in manner of multi-task learning of two tasks: answer-span prediction and context words prediction.

of the word w_i which appears after s_c and that of the word w_i which appears before e_c.

$$p(\text{w}_i \in \mathcal{C}) = p(i \ge s_c) \times p(i \le e_c). \quad (4)$$

Here, we assume the independence between s_c and e_c for computational conciseness. The cumulative distributions of $p(i \ge s_c)$ and $p(i \le e_c)$ are calculated with the following equations:

$$\begin{aligned} p(i \ge s_c) &= \sum_{j \le i} p(j = s_c) \\ p(i \le e_c) &= \sum_{j \ge i} p(j = e_c). \end{aligned} \quad (5)$$

We explicitly force the block attention model to learn context words of a given question by minimizing the cross-entropy of the two probabilities, $p(\text{w}_i \in \mathcal{C})$ and $p_{\text{soft}}(\text{w}_i \in \mathcal{C})$. The loss function for the latent context is defined by the following equation:

$$\begin{aligned} L_{\text{context}} = &- \sum_{1 \le i \le l} p_{\text{soft}}(\text{w}_i \in \mathcal{C}) \log p(\text{w}_i \in \mathcal{C}) \\ &- \sum_{1 \le i \le l} p_{\text{soft}}(\text{w}_i \notin \mathcal{C}) \log p(\text{w}_i \notin \mathcal{C}), \end{aligned} \quad (6)$$

where l is the length of a passage. By averaging L_{context} across all train examples, we get the final context loss function.

3.4 Answer-span Prediction

BLANC predicts answer-span with the context probability, $p(\text{w}_i \in \mathcal{C})$. We use the same answer-span prediction layer as BERT, but we multiply $p(\text{w}_i \in \mathcal{C})$ to the output of the encoder, $\boldsymbol{H}$ to give attention at indices of answer-span within the context, $\mathcal{C}$.

$$\begin{aligned} p(i = s_a) &= \frac{\exp(\mathcal{A}_i \boldsymbol{W_a} \boldsymbol{H}_i + b_s^a)}{\sum_j \exp(\mathcal{A}_j \boldsymbol{W_a} \boldsymbol{H}_j + b_s^a)}, \\ p(i = e_a) &= \frac{\exp(\mathcal{A}_i \boldsymbol{V_a} \boldsymbol{H}_i + b_e^a)}{\sum_j \exp(\mathcal{A}_j \boldsymbol{V_a} \boldsymbol{H}_j + b_e^a)}, \end{aligned} \quad (7)$$

where $\boldsymbol{W_a}$, $\boldsymbol{V_a}$, b_s^a, and b_e^a represent weight and bias parameters for answer-span prediction layer, and $\mathcal{A}_i = p(\text{w}_i \in \text{context})$. The loss function for answer-span prediction is defined by the following equation:

$$\begin{aligned} L_{\text{answer}} = -\frac{1}{2}\{ &\sum_{1 \le i \le l} \mathbb{1}(i = s_a) \log p(i = s_a) \\ &+ \sum_{1 \le i \le l} \mathbb{1}(i = e_a) \log p(i = e_a)\}. \end{aligned} \quad (8)$$

$\mathbb{1}(\text{condition})$ represents an indicator function that returns 1 if the condition is true and returns 0 otherwise. By averaging L_{answer} across all train examples, we get the final answer-span loss function. We define our final loss function as the weighted sum

of the two loss functions:

$$L_{\text{total}} = (1-\lambda)L_{\text{answer}} + \lambda L_{\text{context}}, \quad (9)$$

where λ is a hyper-parameter moderating the ratio of two loss functions.

3.5 Property of Block Attention

$p_{\text{soft}}(\text{w}_i \in \mathcal{C})$ defined at (1) can be represented by the probability distributions calculated by block attention model, $p(\text{w}_i \in \mathcal{C})$. We provide detailed proof in Appendix A.1.

4 Experimental Setup

We validate the efficacy of BLANC on two types of tasks: extractive QA and zero-shot supporting fact prediction. In the extractive QA, we evaluate the overall reading comprehension performance with three QA datasets, and we further analyze the ability of BLANC to discern relevant contexts on passages with multiple answer texts. In zero-shot supporting facts prediction, we train QA models on SQuAD (Rajpurkar et al., 2016) and predict supporting facts (supporting sentences) of answers in HotpotQA (Yang et al., 2018). Due to our experimental computing resource limitation, we compare BLANC to baseline models trained in slightly modified hyperparameter settings instead of the results from their original papers.

4.1 Datasets

SQuAD: SQuAD1.1 (Rajpurkar et al., 2016) is a large reading comprehension dataset for QA. Since the test set for SQuAD1.1 (Rajpurkar et al., 2016) is not publicly available, and their benchmark does not provide an evaluation on the span-based metric, we split train data (90%/10%) into new train/dev dataset and use development dataset as test dataset.

NewsQA & NaturalQ: NewsQA (Trischler et al., 2017) consists of answer-spans to questions generated in a way that reflects realistic information seeking processes in the news domain. NaturalQuestions (Kwiatkowski et al., 2019) is a QA benchmark in a real-world scenario with Google search queries for naturally-occurring questions and passages from Wikipedia for annotating answer-spans. Due to computational limits, we use the curated versions of NewsQA and NaturalQ provided by Fisch et al. (2019). The curated datasets contain train and development set only, so we use the development set as the test set and build new train and dev sets from the train set (90%/10%).

HotpotQA: HotpotQA (Yang et al., 2018) aims to measure complex reasoning performance of QA models and requires finding relevant sentences from the given passages. HotpotQA consists of passages, questions, answer, and corresponding supporting facts (sentences) for each answer. We use the development set in HotpotQA.

4.2 Evaluation Metrics

F1 and EM are evaluation metrics widely used in existing QA models (Rajpurkar et al., 2016). These two metrics measure the number of overlapping tokens between the predicted answers and the ground truth answers. Token matching evaluation treats as correct even answers in unrelated contexts, thus being insufficient to evaluate the context prediction performance. As the alternatives, we propose span-EM and span-F1. We modify the metric proposed in Kwiatkowski et al. (2019) to be suitable for our experiment setting.

Span-F1 and Span-EM: Span-F1 and span-EM are defined with overlapping indices between the predicted span and the ground truth span:

$$\begin{aligned}
\text{Span-P} &= |[s_p, e_p] \cap [s_g, e_g]| / |[s_p, e_p]| \\
\text{Span-R} &= |[s_p, e_p] \cap [s_g, e_g]| / |[s_g, e_g]| \\
\text{Span-F1} &= 2 \times \frac{\text{Span-P} \times \text{Span-R}}{\text{Span-P} + \text{Span-R}} \\
\text{Span-EM} &= \mathbb{1}(s_p = s_g \wedge e_p = e_g)
\end{aligned}$$

Here, s_p / e_p represent the start/end indices of a predicted answer-span in a passage and s_g / e_g denote the start/end indices of the ground truth answer-span in a passage. Span-EM measures exactly matched predicted spans, and Span-F1 quantifies the degree of overlap between the predicted answer-span and the ground truth span.

4.3 Baselines

BERT, RoBERTa, and ALBERT: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2019) are language models built upon the transformer encoder. They use the same model structure, except for the bigger vocabulary size of RoBERTa. Due to the computational limitation, we use 12-layer base models for BERT, and RoBERTa and 24-layer large model for ALBERT.

SpanBERT: SpanBERT (Joshi et al., 2020) has the same model structure and the same parameter

		#Param	Span-F1	Span-EM	F1	EM
NaturalQA	BERT	108M	72.92 ± 0.36	60.63 ± 0.39	76.39 ± 0.26	64.48 ± 0.28
	ALBERT	17M	72.66 ± 0.48	60.31 ± 0.49	75.89 ± 0.36	63.81 ± 0.37
	RoBERTa	124M	75.07 ± 0.17	62.59 ± 0.14	78.54 ± 0.20	66.33 ± 0.09
	SpanBERT	108M	75.16 ± 0.26	62.71 ± 0.37	78.31 ± 0.22	66.60 ± 0.31
	BLANC	108M	**76.99 ± 0.09**	**64.57 ± 0.12**	**80.04 ± 0.06**	**68.33 ± 0.09**
	$\text{SpanBERT}_{\text{large}}$	333M	77.62 ± 0.10	65.28 ± 0.41	80.66 ± 0.11	69.14 ± 0.18
	$\text{BLANC}_{\text{large}}$	333M	**79.04 ± 0.27**	**66.75 ± 0.14**	**81.99 ± 0.16**	**70.59 ± 0.12**
SQuAD1.1	BERT	108M	83.36 ± 0.25	70.74 ± 0.43	88.10 ± 0.14	80.49 ± 0.28
	ALBERT	17M	84.60 ± 0.13	72.04 ± 0.38	88.75 ± 0.20	81.05 ± 0.27
	RoBERTa	124M	85.21 ± 0.25	72.82 ± 0.56	89.91 ± 0.16	82.53 ± 0.44
	SpanBERT	108M	86.67 ± 0.16	74.08 ± 0.13	91.58 ± 0.09	84.97 ± 0.18
	BLANC	108M	**86.89 ± 0.15**	**74.69 ± 0.37**	**91.87 ± 0.13**	**85.30 ± 0.32**
	$\text{SpanBERT}_{\text{large}}$	333M	88.27 ± 0.14	75.96 ± 0.22	93.22 ± 0.08	87.14 ± 0.11
	$\text{BLANC}_{\text{large}}$	333M	**88.42 ± 0.17**	**76.26 ± 0.31**	**93.37 ± 0.05**	**87.30 ± 0.10**
NewsQA	BERT	108M	59.18 ± 0.57	45.53 ± 0.55	65.07 ± 0.52	50.11 ± 0.50
	ALBERT	17M	60.12 ± 0.36	46.54 ± 0.04	66.02 ± 0.35	51.18 ± 0.18
	RoBERTa	124M	61.36 ± 0.63	47.43 ± 0.54	67.28 ± 0.63	52.36 ± 0.64
	SpanBERT	108M	62.26 ± 0.22	48.04 ± 0.48	67.93 ± 0.26	52.85 ± 0.49
	BLANC	108M	**64.39 ± 0.76**	**50.60 ± 0.50**	**70.31 ± 0.66**	**55.52 ± 0.43**
	$\text{SpanBERT}_{\text{large}}$	333M	63.43 ± 0.42	49.03 ± 0.13	69.06 ± 0.55	53.84 ± 0.27
	$\text{BLANC}_{\text{large}}$	333M	**66.48 ± 0.20**	**52.39 ± 0.08**	**72.36 ± 0.01**	**57.40 ± 0.21**

Table 1: Reading comprehension performance of baseline models and BLANC. We conduct experiments on three QA datasets: NaturalQ, SQuAD1.1, and NewsQA. For all evaluation metrics, we report mean and standard deviation of three separate trials. The results show that BLANC outperforms baseline models.

size as BERT. SpanBERT uses span-oriented pre-training for span representation. Since the block attention is stacked on SpanBERT, and to provide detailed results of effectiveness of BLANC, we use both 12-layer SpanBERT-base and 24-layer SpanBERT-large.

4.4 Hyper-parameter Settings

We conduct experiments on limited hyper-parameter settings (e.g. max_len, batch size), as we were limited by computational resources. We use the same hyperparameter settings across all baseline models and BLANC. We set the training batch size to 8, learning-rate to $2 \times e^{-5}$, the number of train epochs to 3, the max sequence length of transformer encoder to 384, warm-up proportion to 10%, and we use the various optimizers used in the respective original papers. We set λ to 0.8, which is the optimal value as we show in Figure 6, for all experiments except the large model experiment on SQuAD1.1. We set $\lambda = 0.2$ in the large model experiment on SQuAD1.1. We use different q, the decreasing ratio in (1), and different window-size for each dataset to reflect the average length of passages of each QA datasets. We set $q = 0.7$ and window-size to 2 on SQuAD which contains relatively short passages, and $q = 0.99$ and window-size to 3 on the other two QA datasets where most passages are longer than SQuAD. q and window-size are optimized empirically.

	Span-F1	Span-EM
RoBERTa	65.99 ± 0.92	60.12 ± 0.86
SpanBERT	63.47 ± 0.72	57.63 ± 0.79
BLANC	**67.07 ± 0.36**	**61.43 ± 0.38**

Table 2: Performance of BLANC on passages of NaturalQ that have answer texts two or more.

5 Results & Discussion

We now present the results for the experiments described in the previous section. We describe the overall reading comprehension performance, highlighting the increased gain for passages with multiple mentions of the answer text. We show that

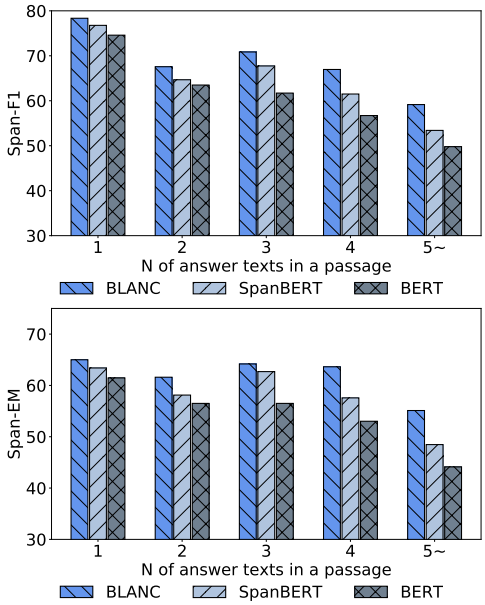


Figure 5: Span-F1 and Span-EM of baseline models and BLANC trained on NaturalQ. We categorize NaturalQ dataset into five groups by number answer texts appeared in a passage: $n = 1, 2, 3, 4$, and $n \geq 5$. BLANC outperforms baseline models on every groups and the performance gap increases as the number of answer texts in a passage increases.

BLANC outperforms other models for zero-shot supporting fact prediction. We also demonstrate the importance of the context prediction loss and the negligible extra parameter and inference time.

5.1 Reading Comprehension

We verify the reading comprehension performance of BLANC with four evaluation metrics (F1, EM, Span-F1, and Span-EM) on three QA datasets: SQuAD, NaturalQ, and NewsQA. We show the results in Table 1 which shows BLANC consistently outperforms all comparison models including RoBERTa and SpanBERT.

We focus on the evaluation metric Span-EM which measures the exact match of the answer-span, and we further highlight the performance gain of BLANC over the most recent SpanBERT model, both base and large. On NaturalQ, BLANC outperforms SpanBERT by 1.86, whereas the performance difference between SpanBERT and RoBERTa is 0.12. On NewsQA, BLANC outperforms by 2.56, whereas the difference between SpanBERT and RoBRTa is 0.61. This pattern holds for the large models as well.

We now compare the performance gain between the datasets. Recall that we showed in Figure 3 the proportion of multi-mentioned answer is smallest in SQuAD, medium for NaturalQ-MRQA, and largest in NewsQA-MRQA. Reading comprehension results show the performance gap of BLANC and SpanBERT increases in the same order, verifying the effectiveness of BLANC on the realistic multi-mentioned datasets.

	Accuracy
BERT	33.34 ± 0.82
$\text{ALBERT}_{\text{large}}$	35.62 ± 1.17
RoBERTa	37.93 ± 0.80
SpanBERT	34.79 ± 0.40
BLANC	$\mathbf{39.80 \pm 1.18}$

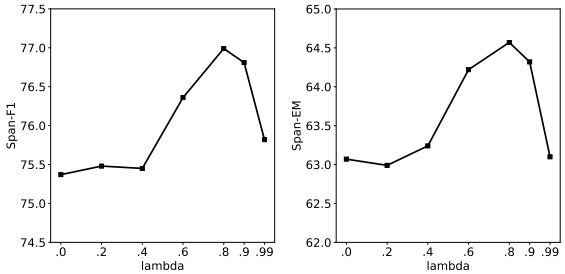
Table 3: Performance on zero-shot supporting fact (supporting sentence) prediction by models trained with SQuAD1.1. BLANC outperforms all other models.

5.2 Performance on Passages with Multi-mentioned Answers

In Section 5.1, we show Span-EM and EM of BLANC and baselines on the entire datasets. However, the context discerning performance is only observed on passages with multiple mentions of the answer text. We investigate the context-aware performance (distinguishing relevant context and irrelevant context) of BLANC by categorizing NaturalQ dataset by the number of occurrences of the answer text in a passage. We subdivide the dataset into five groups: $n = 1, 2, 3, 4$ and $n \geq 5$, where n is the number of occurrences of the answer text in a passage. Figure 5 presents Span-F1 and Span-EM on those subsets of the data. BLANC outperforms SpanBERT and BERT across all subsets, and we show that the performance gain increases as n increases. In Table 2, we explicitly show reading comprehension performance of BLANC on the question-answer pairs of passages with $n \geq 2$ from NaturalQ, and we confirm that block attention method increases context-aware performance of SpanBERT by 3.6 with Span-F1, and by 3.8 with Span-EM, which are larger improvements than the increments on the data including $n = 1$ shown in Table 1.

5.3 Supporting Facts Prediction

We present the results of the zero-shot supporting facts prediction task on HotpotQA dataset (Yang et al., 2018) in Table 3. HotpotQA has ten passages and two supporting facts (sentences) for each question-answer pair. Since HotpotQA has a different data format than the extractive QA datasets, we curate HotpotQA with the following steps. We con-



(a) Span-F1 (b) Span-EM

Figure 6: Analysis on λ for context word prediction for NaturalQ. We adjust λ, weight of (L_{context}), from 0.0 to 0.99 and report Span-F1 and Span-EM. Increasing λ improves answer-span prediction until $\lambda = 0.8$ and then decreases. This decrease is expected as the weight for (L_{answer}) becomes too small.

	Train	Inf
BERT	1.00x	1.00x
$\text{ALBERT}_{\text{large}}$	1.42x	1.89x
RoBERTa	1.01x	1.02x
SpanBERT	1.00x	1.00x
BLANC	1.04x	1.00x

Table 4: Training and inference time of each model measured on the same number of QA pairs.

catenate the ten passages to make one passage. Two supporting facts exist in the passage. By removing each one of them, we build two passages and each of passage contains one supporting fact. We repeat this process for all examples in HotpotQA. As a result, the curated dataset contains triples of one question, one supporting fact, and one passage. We report the accuracy of models by checking if the supporting fact includes the predicted span. We train baseline models and BLANC on SQuAD1.1 and test on the curated development set of HotpotQA dataset. Table 3 shows that BLANC captures sentence relevant to the given question better than other baseline models in zero-shot setting. This result shows that BLANC is capable of applying what it has learned from one dataset to predicting the context of an answer to a question in another dataset.

5.4 Analysis on λ

We verify the relationship between reading comprehension performance and context word prediction task by conducting reading comprehension experiment with $\lambda = [0.2, 0.4, 0.6, 0.8, 0.9, 0.99]$. The hyperparameter λ represents weight of L_{context} in the total loss function L_{total}. Figure 6 shows that the performance increases as λ increases until it reaches 0.8 and decreases after $\lambda = 0.8$. Leveraging the context word prediction task increases reading comprehension performance, and we show efficacy of BLANC. As λ increases, the weight on L_{answer} decreases, so we expect to see a decrease in performance as λ becomes too large.

5.5 Space and Time Complexity

The additional parameters of block attention model come from Eq. (3) in Section 3.3. The number of parameters is $(768 + 1) * 2 = 1538$ when the hidden dimension size of the transformer encoder is 768, and 1538 is negligible considering the total number of parameters in BERT-base (108M). The exact numbers of parameters of baseline models are presented in Table 1. Table 4 shows relative training and inference time of baseline models and BLANC. We measure each model's train time on the same number of train steps and the inference time on the same number of passage-question pairs. Since we use the 24-layer ALBERT-large model which has twice as many layers as other models, ALBERT requires the longest training/inference time, despite its much smaller model size. BLANC requires 4% extra training time which includes the time to generate the soft-labels in (1) and the time to calculate the context word distribution in (4). For inference, BLANC requires negligible additional time on SpanBERT.

6 Conclusion

In this paper, we showed the importance of predicting an answer with the correct context of a given question. We proposed BLANC with two novel ideas: context word prediction task and a block attention method that identifies an answer within the context of a given question. The context words prediction task labels latent context words with the labeled answer-span and is used in a multi-task learning manner. Block attention models the latent context words with negligible extra parameters and training/inference time. We showed that BLANC increases reading comprehension performance, and we verify that the performance gain increases for complex examples (i.e., when the answer occurs two or more times in the passage). Also, we showed the generalizability of BLANC and its context-aware performance with the zero-shot supporting fact prediction task on the HotpotQA dataset.

Acknowledgements

This work was partly supported by NAVER Corp. and Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (2017-0-01780, The technology development for event recognition/relational reasoning and learning knowledge based system for video understanding).

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over Wikipedia graph for question answering. In *ICLR*.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the CNN/Daily mail reading comprehension task. In *ACL*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *ACL*.

Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit, Illia Polosukhin, Alexandre Lacoste, and Jonathan Berant. 2017. Coarse-to-fine question answering for long documents. In *ACL*.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2018. Multi-step retriever-reader interaction for scalable open-domain question answering. In *ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *EMNLP 2019 MRQA Workshop*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NeurIPS*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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. In *ICLR*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *ACL*.

Weikang Li, Wei Li, and Yunfang Wu. 2018. A unified model for document-based question answering based on human-like reading strategy. In *AAAI*.

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. *arXiv*.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019a. A discrete hard EM approach for weakly supervised question answering. In *EMNLP-IJCNLP*.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019b. Multi-hop reading comprehension through question decomposition and rescoring. In *ACL*.

Yilin Niu, Fangkai Jiao, Mantong Zhou, Ting Yao, Jingfang Xu, and Minlie Huang. 2020. A self-training method for machine reading comprehension with soft evidence extraction. *arXiv*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hananneh Hajishirzi. 2017. Bi-directional attention flow for machine comprehension. In *ICLR*.

Swabha Swayamdipta, Ankur P Parikh, and Tom Kwiatkowski. 2018. Multi-mention learning for reading comprehension with neural cascades. In *ICLR*.

Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Densely connected attention propagation for reading comprehension. In *NeurIPS*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Shuohang Wang, Mo Yu, Jing Jiang, Wei Zhang, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro, and Murray Campbell. 2018. Evidence aggregation for answer re-ranking in open-domain question answering. In *ICLR*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*.

A Properties of Block Attention

A.1 Block Attention on a Soft-label

Theorem 1. *There exist two probability distributions, $p(i = s_c)$ and $p(i = e_c)$, that makes $p(w_i \in \mathcal{C})$ equal to $p_{soft}(w_i \in \mathcal{C})$, which is defined as follows:*

$$p_{soft}(w_i \in \mathcal{C}) = \begin{cases} 1.0 & \text{if } i \in [s_a, e_a] \\ q^{|i-s_a|} & \text{if } s_w \le i < s_a \\ q^{|i-e_a|} & \text{if } e_a < i \le e_w \\ 0.0 & \text{if } i < s_w \text{ or } i > e_w \end{cases}. \tag{10}$$

Here, q is the decreasing ratio, which satisfies $q \le 1.0$. s_a and e_a are the start and end indices of an answer-span. s_a and e_a satisfy $s_a \le e_a$. s_w and e_w are the start and end indices of the segments bounded by certain window-size. s_w and e_w satisfy $s_w \le s_a$ and $e_a \le e_w$.

Proof. Based on the independent assumption between s_c and e_c in section 3.3, $p(\mathrm{w}_i \in \mathcal{C})$ becomes multiplication of two probability distributions as follows:

$$p(\mathrm{w}_i \in \mathcal{C}) = p(i \ge s_c) \times p(i \le e_c). \tag{11}$$

Then, the following two cumulative distributions, $p(i \ge s_c)$ and $p(i \le e_c)$, make $p(\mathrm{w}_i \in \mathcal{C})$ equal to $p_{\text{soft}}(\mathrm{w}_i \in \mathcal{C})$:

$$p(i \ge s_c) = \begin{cases} 0.0 & \text{if } i < s_w \\ p_{\text{soft}}(\mathrm{w}_i \in \mathcal{C}) & \text{if } s_w \le i < s_a \\ 1.0 & \text{if } s_a \le i \end{cases}, \tag{12}$$

$$p(i \le e_c) = \begin{cases} 1.0 & \text{if } i \le e_a \\ p_{\text{soft}}(\mathrm{w}_i \in \mathcal{C}) & \text{if } e_a < i \le e_w \\ 0.0 & \text{if } e_w < i \end{cases}. \tag{13}$$

Since block attention method can predict any form of $p(i = s_c)$ and $p(i = e_c)$, any soft-label can be represented by block attention method. □

A.2 Block Attention on Multiple Spans

Block attention model can be expanded to predict multiple spans.

Theorem 2. *Any form of the following $p_{multi\text{-}span}(w_i \in \mathcal{C})$, which has m-blocks, can be represented by the multiplication of a scaling factor, k, and the probability distribution calculated by block attention model, $p(w_i \in \mathcal{C})$.*

$$p_{multi\text{-}span}(w_i \in \mathcal{C}) = \begin{cases} a & \text{if } i \in \mathcal{B}_1 \vee ... \vee i \in \mathcal{B}_m \\ \epsilon & \text{otherwise} \end{cases}. \tag{14}$$

Here, $\mathcal{B}_i$ is the set of indices of the i-th span, $\mathcal{B}_i = [s_i^b, e_i^b]$. s_i^b and e_i^b are the start and end indices of $\mathcal{B}_i$. $\mathcal{B}_i$ satisfies $s_i^b \le e_i^b$ and $e_i^b < s_{i+1}^b$ for all i.

Proof. Following two cumulative distributions and the scaling factor make $k \times p(i \ge s_c) \times p(i \le e_c)$ equal to $p_{\text{soft}}(\mathrm{w}_i \in \mathcal{C})$ for all i.

$$p(i \ge s_c) = \begin{cases} (\frac{\epsilon}{a})^m & \text{if } i < s_1^b \\ (\frac{\epsilon}{a})^{m-j} & \text{if } s_j^b \le i < s_{j+1}^b; j \in [1, m) \\ 1.0 & \text{if } s_m^b \le i \end{cases} \tag{15}$$

$$p(i \le e_c) = \begin{cases} 1.0 & \text{if } i \le e_1^b \\ (\frac{\epsilon}{a})^j & \text{if } e_j^b < i \le e_{j+1}^b; j \in [1, m) \\ (\frac{\epsilon}{a})^m & \text{if } i > e_m^b \end{cases} \tag{16}$$

$$k = \epsilon \Big(\frac{a}{\epsilon}\Big)^m \tag{17}$$

Since block attention model can predict any form of $p(i = s_c)$ and $p(i = e_c)$, $p_{\text{multi-span}}(w_i \in \mathcal{C})$ can be represented by the multiplication of a scaling factor and the probability distribution calculated by block attention model. □

B Semantic Similarity Between Context Words and Questions

Soft-labeling method assumes that words near an answer-span are likely to be included in the context of a given question. We provide the basis of this assumption with the question-word similarity experiment. The question-word similarity is calculated with the cosine similarity between word vectors and question vectors. We use word2vec vectors and calculate the question vectors by averaging word vectors in the questions. Figure 7 shows that words adjacent to the answer-spans have the most similar

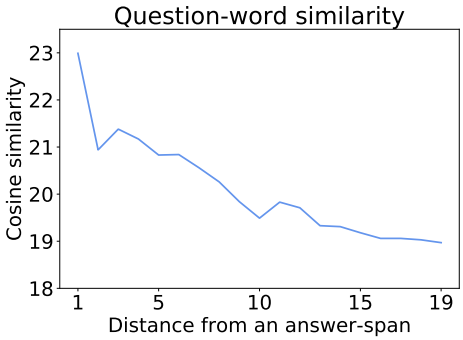


Figure 7: The semantic similarity between a given question and words in a passage. The x-axis represents the distance between a word and an answer-span. The y-axis represents the cosine similarity between the question and the word on 100 scale. Words near an answer-span are likely to have a similar meaning to a given question.

WS	Span-F1	Span-EM	F1	EM	AVG
1	77.06	64.52	80.02	68.04	72.41
2	75.95	63.35	79.80	67.84	71.73
3	76.99	64.41	80.05	68.38	72.45
4	76.38	63.96	80.09	68.44	72.21
5	77.01	64.33	80.02	68.04	72.35
7	76.37	64.19	79.81	68.11	72.12
21	76.65	64.14	79.96	68.06	72.20

Table 5: The performance of BLANC on NaturalQuestions. We vary window-size to find the optimal context size. AVG represents the average of the four performances.

meaning to given questions. Also, the similarity decreases as the distance between the words and the answer-spans increases. From the results, we verify the assumption.

C Details about Hyperparameter Settings

We vary window-size, and λ to find the optimal hyperparameters of BLANC.

C.1 Analysis on Window-size

Table 5 shows the performance of BLANC trained on NaturalQuestions with window-size $= [1, 2, 3, 4, 5, 7, 21]$. AVG represents the average of the four performances. BLANC shows the best AVG performance at WS $= 3$, and we set window-size to 3 for NaturalQuestions and NewsQA experiments.

λ	Span-F1	Span-EM
0.2	88.42 ± 0.17	76.26 ± 0.31
0.8	88.30 ± 0.16	75.71 ± 0.30

Table 6: The performance of BLANC on SQuAD1.1 with two different λ settings.

C.2 Varying λ on SQuAD1.1

Table 6 shows the performance of BLANC with two different λ settings on SQuAD1.1. The results show that BLANC performs better at $\lambda = 0.2$ than $\lambda = 0.8$ (the optimal value for NaturalQuestions) on SQuAD1.1. We set λ to 0.2 in SQuAD1.1 experiments.