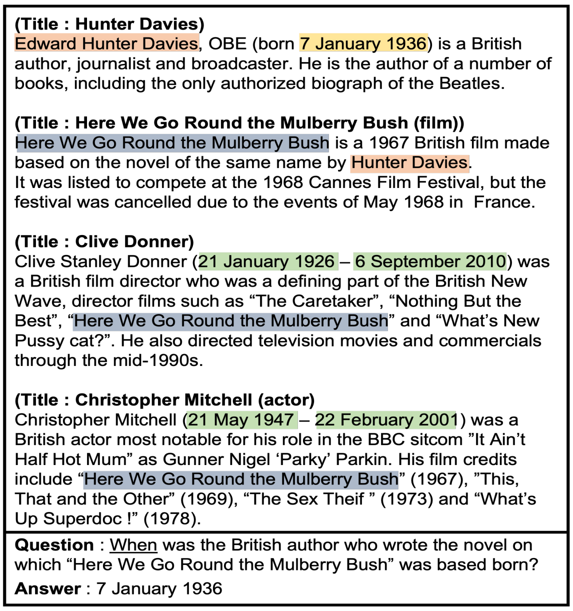
Enhancing text comprehension for Question Answering with Contrastive Learning

Seungyeon Lee[1], **Minho Lee**[1,2,3,*]
[1]Department of Artificial Intelligence, Kyungpook National University
[2]School of Electronics Engineering, Kyungpook National University
[3]ALI Co., Ltd.
statai3237@knu.ac.kr, mholee@knu.ac.kr

Abstract

Although Question Answering (QA) have advanced to the human-level language skills in NLP tasks, there is still a problem: the QA model gets confused when there are similar sentences or paragraphs. Existing studies focus on enhancing the text understanding of the candidate answers to improve the overall performance of the QA models. However, since these methods focus on re-ranking queries or candidate answers, they fail to resolve the confusion when many generated answers are similar to the expected answer. To address these issues, we propose a novel contrastive learning framework called ContrastiveQA that alleviates the confusion problem in answer extraction. We propose a supervised method where we generate positive and negative samples from the candidate answers and the given answer, respectively. We thus introduce ContrastiveQA, which uses contrastive learning with sampling data to reduce incorrect answers. Experimental results on four QA benchmarks show the effectiveness of the proposed method.



(Title : Hunter Davies)
Edward Hunter Davies, OBE (born 7 January 1936) is a British author, journalist and broadcaster. He is the author of a number of books, including the only authorized biograph of the Beatles.

(Title : Here We Go Round the Mulberry Bush (film))
Here We Go Round the Mulberry Bush is a 1967 British film made based on the novel of the same name by Hunter Davies.
It was listed to compete at the 1968 Cannes Film Festival, but the festival was cancelled due to the events of May 1968 in France.

(Title : Clive Donner)
Clive Stanley Donner (21 January 1926 – 6 September 2010) was a British film director who was a defining part of the British New Wave, director films such as "The Caretaker", "Nothing But the Best", "Here We Go Round the Mulberry Bush" and "What's New Pussy cat?". He also directed television movies and commercials through the mid-1990s.

(Title : Christopher Mitchell (actor))
Christopher Mitchell (21 May 1947 – 22 February 2001) was a British actor most notable for his role in the BBC sitcom "It Ain't Half Hot Mum" as Gunner Nigel 'Parky' Parkin. His film credits include "Here We Go Round the Mulberry Bush" (1967), "This, That and the Other" (1969), "The Sex Theif " (1973) and "What's Up Superdoc !" (1978).

Question : When was the British author who wrote the novel on which "Here We Go Round the Mulberry Bush" was based born?
Answer : 7 January 1936

Figure 1: An example of HotpotQA dataset that consists of several paragraphs in one context. HotpotQA consists of two supporting paragraphs and eight distractor settings. A distractor paragraph appears to be related to the question but is used to induce confusion.

1 Introduction

Large language model (LLM) is developing rapidly, and ChatGPT (Ouyang et al., 2022), which can perform almost all natural language processing tasks, is receiving tremendous attention. ChatGPT can be easily used by anyone through a conversation interface, and ChatGPT responds appropriately to the user's query. It is clear that ChatGPT has led to tremendous development and success, but it has a fatal problem called hallucination (Alkaissi and McFarlane, 2023; Bang et al., 2023). A generative model such as ChatGPT can have serious consequences if poor decisions are made when it is used in specific domains such as finance, law, and medicine (Shahriar and Hayawi, 2023). To avoid this hallucination problem in specific domains, finding the correct answer within a given document is more reliable. In this paper, we focus on extractive question answering (QA), which extracts the correct answer to a given question in a context. In the field of natural language processing, there are several attempts to solve extractive QA problems that can increase the ability to answer and reason like humans, such as HotpotQA dataset (Yang et al., 2018; Gao et al., 2021; Saxena et al., 2020). However, it is still challenging for the extractive QA model to have a human-level understanding or comprehension skills. In general, humans can easily find and select the correct answer among many similar answers related to the given question. On the contrary, the current QA model is somewhat less capable of selecting the correct answer if the correct answer is confused within the context (Gupta et al., 2018; Kratzwald et al.,

[*]Corresponding author

Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023), pages 75–86
July 13, 2023 ©2023 Association for Computational Linguistics

2019; Mao et al., 2021; Glass et al., 2022). To deal with these issues, the conventional methods for improving QA performance use answer re-ranking methods to find the correct answers (Iyer et al., 2021; Majumder et al., 2021). In (Iyer et al., 2021), they use a cross-attention method to build an answer re-ranking model. The (Majumder et al., 2021) seeks to improve the performance of the QA model by explicitly assigning scores to candidate sentences according to the overlap of named entities in the question and sentence. Although the answer re-ranking methods are simple and intuitive, they employ methods of re-ranking answers over a smaller set of candidates rather than considering confusing cases of answer selection. As shown in Fig. 1, a typical QA model explores each sentence to extract an answer to a question. For complex cases like Fig. 1, the QA model may confuse selecting a correct answer during answer extraction and extracting other similar answers. These cases are mainly because the distractor words confuse finding the correct answer.

With this motivation, we propose a novel QA model called ContrastiveQA that uses contrastive learning to pull semantically similar sentences closer and push different sentences further. In addition, the proposed model introduces a method of positive and negative sampling with candidate answers. ContrastiveQA performs the contrastive learning by generating positive and negative samples with the given answer and candidate answers obtained from a pre-finetuned QA model. The proposed model extracts an answer close to the correct answer because the QA model uses a contrastive objective to alleviate the confusing problem of selecting the correct answer.

We demonstrate the effectiveness of our approach on four extractive QA datasets: HotpotQA (Yang et al., 2018), SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017) and NewsQA (Trischler et al., 2017). We obtain notable performance for the proposed ContrastiveQA, a method learned by contrastive learning, on extractive QA datasets. In particular, the experimental results demonstrate that the proposed method performs significantly better than the previous approaches due to the positive and negative sampling.

Contributions. The main contributions of the proposed ContrastiveQA are as follows :

- We propose a novel method called, ContrastiveQA that effectively solves the problem of ambiguous answer selection and confusion.
- We introduce positive and negative sampling to perform contrastive learning better using QA datasets.
- Experimental results show that the proposed models outperform the baseline models on the extractive QA datasets.

2 Related Works

Contrastive learning for Question Answering. Contrastive learning has achieved remarkable performance in supervised/unsupervised learning and is actively used in image processing (Liang et al., 2020; Wang et al., 2021, 2022). The main goal of contrastive learning is to learn representations that keep similar samples close together and dissimilar samples far apart. In the field of natural language processing, contrastive learning is used to train text representation. Contrastive Domain Adaptive For Question Answering (CAQA) (Yue et al., 2021) proposes a framework for answering out-of-domain questions combining question generation and contrastive domain adaptation to learn domain invariants. This setup can transfer knowledge to the target domain without additional training that can effectively answer the out-of-domain questions. Cross Momentum Contrastive Learning (xMoCo) (Yang et al., 2021) used fast and slow encoders to learn and optimize phrase-question and question-phrase matching. They demonstrated its effectiveness in open-domain QA similar to MoCo (He et al., 2020). It effectively retained large speech samples requiring different question and phrase encoders. (Hu et al., 2022) proposed Momentum Contrastive Pre-Training for Question AnSwering (MCROSS) that employed contrastive learning to align knowledge learned from cloze-type samples to answer natural language questions. This momentum contrastive learning method improved the performance of pre-trained models when answering questions by maximizing the consistency of the distribution of predicted answers. It showed remarkable performance in QA tasks for supervised and zero-shot settings. (Caciularu et al., 2022) applied contrastive learning to reinforce the similarity between the question and evidence sentences. They introduced information encoding to relate the questions to the sentences in order to optimize questions.

They also used similarity mechanisms in specific subspaces by linear projections of raw representations. This method showed consistent improvement in benchmarks with long document context (Yang et al., 2018; Dasigi et al., 2021).

Unlike previous studies, our proposed approach aims to avoid ambiguous answer extraction by processing candidates similar to the correct answer in a set of candidates using contrastive learning.

3 Proposed model

In this work, we propose a novel question-answering (QA) model, called ContrastiveQA. The proposed model allows accurate answers to be extracted by adding a contrastive loss term, which alleviates the confusion and errors that may occur when using the existing QA model.

To achieve this purpose, we divide the training phase into three main steps: candidate answer extraction, positive and negative sampling, and ContrastiveQA to perform contrastive learning. In particular, we introduce a method for creating data to perform ContrastiveQA, where the correct answer is known, in the positive and negative sampling step.

The overall flow of the proposed ContrastiveQA model is shown in Fig. 2.

3.1 Candidate answer extraction

We use the pre-finetuned QA model to perform candidate answer extraction [1]. Like the existing QA models, inputs are questions and contexts, and outputs are answers. When extracting answers, we do not extract only one answer but multiple answers in the order of high confidence score. The QA model gives a confidence score for each of the output answers. And, from among the multiple answers obtained, the most optimal number of candidate answers required for learning are heuristically obtained through experiments (see as subsection 5.2).

Such an extracted candidate answer has a start position and an end position that indicates the answer location within a given context. We represent the candidate answer set by $\mathbb{A}$ that has $\mathbb{A}^{start}$ and $\mathbb{A}^{end}$ in Eq. (1) :

$$\mathbb{A}_{ik}(\mathbb{A}_{ik}^{start}, \mathbb{A}_{ik}^{end}) = QA(q_i, c_i) \qquad (1)$$

In $\mathbb{A}_{ik}$, i is the number of contexts and k is the number of candidate answer sets.

3.2 Positive and Negative Sampling

In this subsection, we describe the process of obtaining positive and negative samples to perform contrastive learning. In general, dropout, masking, random change, and back-translation are used to obtain positive and negative samples for contrastive learning in NLP tasks. These methods can be easily applied to obtain samples, but it is challenging to apply to QA tasks because finding correct answers to questions is crucial. We thus utilize the answer's span (start, end) position information to reconstruct the data.

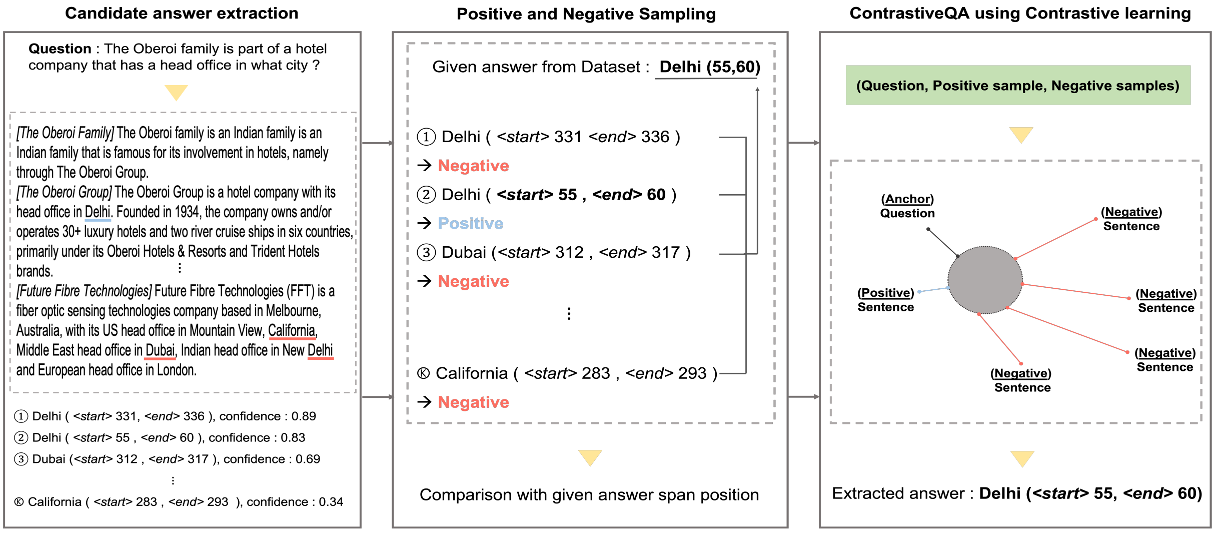
We refer to positive and negative samples as the set of extracted candidates to mitigate the confusing problem of QA models in subsection 3.1. We distinguish between positive and negative samples by comparing the given answers in the training data with the candidate answers.

Algorithm 1 Procedure of positive and negative sampling for Supervised ContrastiveQA

```
Input: Candidate answer ($\mathbb{A}_{ik}$),
       Given answer ($A_i$)
Output: Positive, Negative samples
1: Compare the span positions of the candidate answer and the given answer.
2: Given answer (start, end) position : ($A_i^s, A_i^e$)
3: Candidate answer (start, end) position : ($\mathbb{A}_{ik}^s, \mathbb{A}_{ik}^e$)
4: if ($A_i^s, A_i^e$) = ($\mathbb{A}_{ik}^s, \mathbb{A}_{ik}^e$) then
5:     Positive sample
6: else if $A_i = \mathbb{A}_{ik}$ and ($A_i^s, A_i^e$) $\neq$ ($\mathbb{A}_{ik}^s, \mathbb{A}_{ik}^e$) then
7:     Negative sample
8: end if
9: return Positive, Negative samples for each context
```

We consider a candidate answer with the same token as a correct answer token and the same span position (start, end) as a positive sample. We consider a positive sample when the candidate answer and the actual correct answer have the same token positions. Conversely, if the candidate answer and the actual correct answer do not have the same token position, it is considered a negative sample.

[1] The QA model used to extract the candidate answer is Longformer (Beltagy et al., 2020), which is the backbone model in the experiments. To train the pre-finetuned QA model, we use the data except for the unanswerable of SQuAD v2.0 (Rajpurkar et al., 2018) for training to avoid overlapping data contents. Here, we use the answerable context-question-answer pair of SQuAD v2.0.


Candidate answer extraction
Positive and Negative Sampling
ContrastiveQA using Contrastive learning

Question : The Oberoi family is part of a hotel company that has a head office in what city ?

[The Oberoi Family] The Oberoi family is an Indian family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group.
[The Oberoi Group] The Oberoi Group is a hotel company with its head office in Delhi. Founded in 1934, the company owns and/or operates 30+ luxury hotels and two river cruise ships in six countries, primarily under its Oberoi Hotels & Resorts and Trident Hotels brands.
⋮
[Future Fibre Technologies] Future Fibre Technologies (FFT) is a fiber optic sensing technologies company based in Melbourne, Australia, with its US head office in Mountain View, California, Middle East head office in Dubai, Indian head office in New Delhi and European head office in London.

① Delhi (<start> 331, <end> 336), confidence : 0.89
② Delhi (<start> 55 , <end> 60), confidence : 0.83
③ Dubai (<start> 312 , <end> 317), confidence : 0.69
⋮
Ⓚ California (<start> 283 , <end> 293), confidence : 0.34
Given answer from Dataset : Delhi (55,60)
① Delhi (<start> 331 <end> 336)
→ Negative
② Delhi (<start> 55 , <end> 60)
→ Positive
③ Dubai (<start> 312 , <end> 317)
→ Negative
⋮
Ⓚ California (<start> 283 , <end> 293)
→ Negative
Comparison with given answer span position
(Question, Positive sample, Negative samples)
(Anchor) Question
(Negative) Sentence
(Positive) Sentence
(Negative) Sentence
(Negative) Sentence
(Negative) Sentence
Extracted answer : Delhi (<start> 55, <end> 60)

Figure 2: The overall flow of our ContrastiveQA model using the example of HotpotQA. (Left): The process of extracting candidate answers through a pre-finetuned QA model. (Center): In the positive and negative sampling step, positive and negative samples are designated using the given answer information. (Right): Contrastive learning is performed using positive and negative sampling data.

To utilize the positive and negative samples created in long sequences, the sentences with each sample are also designated as candidate sentences corresponding to the candidate answers. We define a candidate sentence set as Eq. (2) :

$$\mathbb{S}_{ik} = \{s_{ij} | \mathbb{A}_{ik} \subset s_{ij}, 1 \leq j \leq M\} \qquad (2)$$

where the i-th context has M sentences, and $\mathbb{S}_{ik}$ indicates the sentence to which $\mathbb{A}_{ik}$ belongs. In Algorithm 1, we explain the overall flow for the positive and negative sampling phase.

3.3 ContrastiveQA using contrastive learning

We attempt to address QA problems that can be misleading for distractor words or similar answers using positive and negative samples. First, we use Longformer (Beltagy et al., 2020) as a backbone model to extract an answer with a question and context. We use the context consisting of 15 candidate sentences rather than using the given context. We detail the experiments and analysis of context construction in subsection 5.2. As in previous studies (Beltagy et al., 2020; Zaheer et al., 2020; Caciularu et al., 2021, 2022), we use the special tokens <q> and </s> to indicate question and sentence boundaries to represent sentences in a long context or document as one long sequence. We configure the ContrastiveQA model's input as $[CLS]$ <q> $q_1, q_2, q_3, \cdots$ </s> $s_{11}, s_{12}, \cdots$ </s> $s_{21}, s_{22}, \cdots$ </s> $s_{n1}, s_{n2}, \cdots$ by adding a special token like the input format introduced in Longformer (Beltagy et al., 2020). In the input format, q_i represents a question token, and s_{jk} represents the k^{th} token of the j^{th} sentence. The special tokens are added to the vocabulary and initialized before fine-tuning the model. We define L_{CE} using the cross-entropy loss and L_{QC} using the InfoNCE loss (Oord et al., 2018) to perform contrastive objectives. Then, we define the final loss, L_{CL}, by adjusting L_{CE} and L_{QC} with λ. In the final loss (L_{CL}), λ is weight hyperparameter. We heuristically adjust the λ value to 0.6.

$$\mathcal{L}_{CL} = \lambda \mathcal{L}_{CE} + (1 - \lambda)\mathcal{L}_{QC} \qquad (3)$$

L_{CE} calculates the difference between the probability distribution of y and $\hat{y}$, and during training, it learns to reduce this difference. We define it as Eq. (4) :

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \qquad (4)$$

In L_{QC}, positive sentences (s^+) and question (q) representations are pulled to be close to each other, while negative samples (s^-) are pushed to be far apart from each other. We define it as Eq. (5) :

$$\mathcal{L}_{QC_i} = -log \frac{e^{sim(q_i,s^+)/\tau}}{e^{sim(q_i,s^+)/\tau} + \sum_j^{k-1} e^{sim(q_i,s_j^-)/\tau}} \qquad (5)$$

Dataset	Task	QA type	Source	Train set
HotpotQA	Multi-hop question answering	Extractive	Wikipedia	90.4k
SQuAD	Question answering	Extractive	Wikipedia	87.5k
TriviaQA	Reading Comprehension	Extractive	Wikipedia, Web	110k
NewsQA	Reading Comprehension	Extractive	CNN news article	92.5k

Table 1: HotpotQA, SQuAD, TriviaQA, and NewsQA datasets used for experiments.

where sim(·,·) represents similarity between two elements and τ is a configurable temperature parameter. We use Sentence-BERT (SBERT)[2] to calculate the similarity between two sentences. In detail, we experiment with the similarity function in subsection 5.4 and discuss it. Formally, we define the contrastive loss as the sum of the negative log-likelihood losses over all input examples, as introduced in (Caciularu et al., 2022). Here, each loss term distinguishes between positive and negative samples. L_{QC_i} serves one example each, and the final loss (L_{QC}) is obtained as the average loss value over all the examples.

4 Experiment Setup

In this section, we provide experimental datasets and detail about the experiments. We summarize the number of train datasets and the source of the dataset in Table 1.

4.1 Datasets

We use HotpotQA (Yang et al., 2018), SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017), and NewsQA (Trischler et al., 2017) validation datasets for evaluation.

HotpotQA (Yang et al., 2018) is an multi-hop QA dataset that requires reasoning. In this experiment, we use the HotpotQA-distractor setting. HotpotQA includes yes and no as the answer type, so we add 'yes no' character at the beginning of the context.

SQuAD (Rajpurkar et al., 2016) is a set of QA pairs derived from Wikipedia articles. In SQuAD, the correct answer can be found as a sequence of tokens within the given context. This dataset consists of over 100,000 question-answer pairs.

TriviaQA (Joshi et al., 2017) is a realistic text-based QA dataset obtained from Wikipedia and the Web. TriviaQA has relatively complex and structured questions, with considerable syntactic and lexical variability.

NewsQA (Trischler et al., 2017) is a reading comprehension dataset for CNN news articles. Some answers to questions in the NewsQA dataset may sometimes require reasoning.

4.2 Training details

In the experiment, we evaluate the ContrastiveQA to alleviate the problem of confusion in answer extraction. We use the Longformer-base model of the Huggingface Transformer as the backbone model[3] (Wolf et al., 2020). We use four RTX 8000 GPUs in our experiments. We experiment with a batch size of 8 for the Longformer-based model and 32 for the BERT-based model. We conduct experiments in different optimal experimental settings for each dataset. The learning rate for the datasets HotpotQA and TriviaQA is set to $3e^{-5}$, and the model is trained for five epochs. For other two datasets, SQuAD and NewsQA, the learning rates are set to $5e^{-5}$ and $1e^{-4}$, respectively, and the model is trained for three epochs and five epochs, respectively.

4.3 Baselines

We compare the ContrastiveQA model with eight baselines below :

- **Long-context QA Models**

  - Longformer (Beltagy et al., 2020) encodes long inputs using both sliding-window local attention and global attention.

  - CDLM (Caciularu et al., 2021) is a cross-document language model that learns relationships between documents by pre-training a set of related documents.

  - BigBird (Zaheer et al., 2020) uses up to 4,096 tokens by applying block sparse attention to efficiently look at the tokens.

- **Extractive QA Models**

[2] SBERT(Reimers and Gurevych, 2019) is modified BERT that uses siamese and triplet network structure to derive semantically meaningful sentence embeddings. The code is available at `https://www.sbert.net/`.

[3] `https://huggingface.co/`

	Supervised Setting		
Model	**Ans**	**Sup**	**Joint**
Longformer	76.4	82.2	69.3
CDLM	77.2	85.4	71.7
BigBird	76.8	83.1	70.2
BERT	69.7	74.3	62.1
RoBERTa	73.1	79.2	65.0
SpanBERT	71.5	78.6	63.4
SLED	75.8	80.4	67.9
HGN	81.3	86.2	72.5
ContrastiveQA (ours)	**83.9**	**89.4**	**75.6**

Table 2: Experimental results on the HotpotQA dataset. The best performance is indicated in bold. In this table, **Ans** refers to the answer span, and **Sup** refers to the results of supporting facts. **Joint** is the harmonic mean value obtained by multiplying the recall and precision of **Ans** and **Sup**. The metric for evaluation is the F1 score.

- BERT (Devlin et al., 2019) is a pre-trained language model that can understand the context by pre-training MLM (masked language model) and NSP (next sentence prediction).

- RoBERTa (Liu et al., 2019) is a Robustly Optimized BERT Pretraining Approach, and it shows better performance than the existing BERT by adjusting the hyperparameters and training data size of BERT.

- SpanBERT (Joshi et al., 2020) is a model that masks random spans unlike BERT, which masks random tokens and learns a span boundary representation to represent and predict text spans well.

- **Previous Multi-hop QA Models**

- HGN[4] (Fang et al., 2020) is a hierarchical graph network that reinforces answer/evidence prediction through graphs subdivided step by step in a hierarchical framework.

- SLED[5] (Ivgi et al., 2023) is a method of processing long texts using a fusion-in-decoder-based encoder-decoder language model.

5 Model Analysis

5.1 Extractive QA Results

We compare previously superior baseline models for extractive QA with ContrastiveQA.

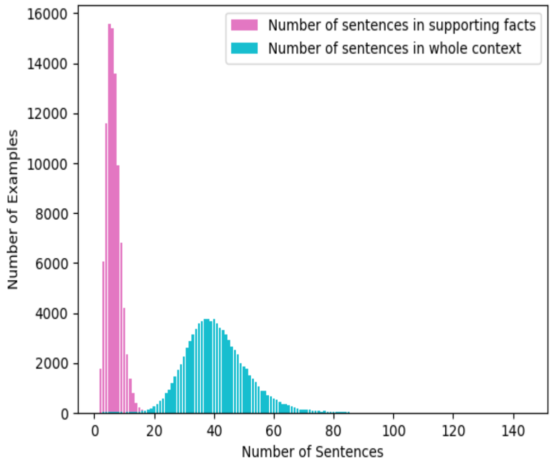


Figure 3: Distribution of the number of sentences in a given context and the number of sentences in supporting facts on the HotpotQA dataset.

Supervised ContrastiveQA. We evaluate the performance of models using the F1-score for four benchmarks. Table 2 shows the results of HotpotQA trained by the supervised learning method. Our proposed model results demonstrate improved performance compared to the baseline. In particular, HopotQA, which we used in this experiment, is based on a distractor setting. HotpotQA contains distractor words within the context, causing more confusion in QA tasks. The proposed ContrastiveQA model improves performance by avoiding confusion. In addition, we show excellent performance compared to the existing baseline model based on F1-score for the SQuAD, TriviaQA, and NewsQA datasets in Table 3. Those results demonstrate the effectiveness of positive and negative sampling and contrastive objectives to create positive and negative samples to avoid confusion in a supervised setting.

5.2 Effect of Selecting Candidate sets

We conduct an ablation study with experimental data to prove the effectiveness of the components of the proposed model, ContrastiveQA. We perform an ablation experiment to find the optimal number of candidates in candidate answer extraction introduced in section 3.1. As shown in Fig. 3, the HotpotQA data of the distractor setting has two gold paragraphs and eight distractor paragraphs, so the number of total sentences is very large. However, in practice, the number of sentences required to infer the correct answer is significantly less than

[4] https://github.com/yuwfan/HGN

[5] https://github.com/Mivg/SLED

Model	Supervised Setting		
	SQuAD	TriviaQA	NewsQA
Longformer	91.3	75.8	72.4
CDLM	91.7	76.9	71.8
BigBird	92.0	77.4	73.7
BERT	87.6	70.9	66.1
RoBERTa	89.0	72.2	68.6
SpanBERT	90.2	73.5	71.2
ContrastiveQA (ours)	**93.8**	**81.7**	**77.8**

Table 3: Experimental results on the SQUAD, TriviaQA, and NewsQA dataset. The best performances are indicated in bold. The metric for evaluation is the F1 score.

Model	**k**	**F1-score**
	5	69.8
	10	72.3
Longformer	**15**	**76.4**
	20	75.6
	All	73.9
	5	70.4
	10	74.2
CDLM	**15**	**77.2**
	20	76.5
	All	76.0
	5	79.5
	10	82.6
ContrastiveQA (ours)	**15**	**83.9**
	20	82.3
	All	81.8

Table 4: Selection of the optimal number of candidate sets (k) on the HotpotQA validation set. Comparison of performance among Longformer, CDLM, and ContrastiveQA, where k is set to 5, 10, 15, 20, and all data. The best performance is in bold.

the total number of sentences. For this reason, we adopt a method of extracting and using sentences related to the correct answer rather than considering the entire sentence. We experiment using the HotpotQA validation set that has the longest sequence with ten paragraphs as a representative example. We experiment by setting k, the number of candidate sets, to 5, 10, 15, 20, and all data. One becomes a positive sample from the total number of candidates, and the rest becomes negative samples. As shown in Table 4, we get the best performance when the number of candidates is 15 in all three models. Since HotpotQA data are complex and require reasoning, the performance may be reduced if the candidate set is too small. Moreover, if the entire dataset is used up, performance may be deteriorated because the sequence length exceeds the capacity (e.g., 512, 1024, 4096) the model can accommodate. We obtain the optimal k of 15 in the ablation experiment. This means we set one positive sample and the rest as negative samples, 15 samples in the experiment.

λ	HotpotQA (F1)
0.4	79.1
0.5 (Standard)	81.6
0.6 (Optimal)	83.9
0.7	80.7
1.0 (Cross-Entropy loss)	76.4

Table 5: Experimental results on the effect of contrastive loss. When λ is 1.0, it is the same case as using only the cross-entropy loss.

5.3 Effect of Contrastive Loss

In this subsection, we analyze the effect of contrastive loss, a key component of our proposed model. Through experiments, we derive the most optimal result when λ is 0.6. First, we set the weight as 0.5 between the cross-entropy and contrastive loss to find the optimal λ value. Then, we experiment with 0.4 and 0.6 and confirm that the performance is lower when the weight is 0.4 than when the weight is 0.5, but the performance is better when the weight is 0.6. As shown in Table 5, when λ is 1, this is the result of using only the cross-entropy loss, and we demonstrate that the QA result is improved by using the contrastive loss through this ablation experiment. In Table 6, we show an example of both when the contrastive loss is applied and when it is not to demonstrate the effect of contrastive loss more intuitively.

Data fields	**Example**
Question	Where is the stadium at which 1964 Georgia Tech Yellow Jacket football team played their home game located?
Context	The 1964 Georgia Tech Yellow Jackets football team represented the Georgia Institute of Technology during the 1964 college football season. The Yellow Jackets were led by 20th-year head coach Bobby Dodd, and played their home games at Grant Field in Atlanta, Georgia. They competed as independents for the first time since 1920, after dropping from the Southeastern Conference in 1963. Bobby Dodd Stadium at Historic Grant Field is the football stadium located at the corner of North Avenue at Techwood Drive on the campus of the Georgia Institute of Technology in Atlanta. It has been home to the Georgia Tech Yellow Jackets football team, often referred to as the "RamblinẂreck", in rudimentary form since 1905 and as a complete stadium since 1913. · · · .
Expected Answer (Label)	North Avenue at Techwood Drive
Answer (Longformer)	Grant Field (confidence score : 0.4751)
Answer (ContrastiveQA)	North Avenue at Techwood Drive (confidence score: 0.6833)

Table 6: Example results of Longformer and ContrastiveQA for the HotpotQA validation set. QA results for Longformer and ContrastiveQA are indicated in pink and cyan, respectively.

Similarity	HotpotQA		
	Ans	Sup	Joint
Cosine-similarity	83.5	89.1	74.3
SBERT	83.9	89.4	75.6

Table 7: Experimental results for the similarity function of the contrastive loss. HotpotQA validation set results (F1) for cosine-similarity and SentenceBERT (SBERT).

5.4 Effect of Similarity Function

As briefly described in subsection 3.3, we perform experiments on the similarity function of the contrastive loss term. We experiment using SBERT (Reimers and Gurevych, 2019), which applies not only cosine-similarity, commonly used to calculate the similarity of two sentences (question, sentence), but also sentence-level embedding. In Table 7, we show the results obtained using cosine-similarity and SBERT on the HotpotQA validation set. As a result of HotpotQA, we demonstrate that SBERT performs better than cosine similarity in all three items: answer span, supporting facts, and joint F1.

6 Conclusion

In this work, we proposed a novel ContrastiveQA to alleviate the problem of confusion in answer extraction using contrastive learning. The proposed model ContrastiveQA consists of three tasks: 1) candidate answer extraction, 2) positive and negative sampling, 3) ContrastiveQA using contrastive learning. We demonstrated the effectiveness of the proposed model by outperforming the performance of the baseline models on four benchmark datasets. In the future, we would like to adopt a method to specify the number of samples required in contrastive learning dynamically. This approach created an adaptive mechanism to use the number of samples for each input. This method could improve the overall performance of the model.

7 Limitations

We obtain notable QA performance through experiments. However, we conduct many experiments to find the optimal candidate for ContrastiveQA. Many of these experiments inevitably consume a lot of time and energy, and we have to heuristically determine the number of candidate sets through experiments in a limited environment. We intend to alleviate the current problems by adding a module that can solve these problems in our future research. For example, the Longformer model takes almost a day to process long text for each epoch to train. Therefore, we use smaller batch sizes with a limited number of GPUs to train the LongFormer model. However, due to the lack of GPU resources, the optimal weight of the proposed framework cannot be learned. Thus, it is necessary for further research on model weight reduction to mitigate computational resource problems.

Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-02068, Artificial Intelligence Innovation Hub), and Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD230017TD).

References

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2).

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Avi Caciularu, Ido Dagan, Jacob Goldberger, and Arman Cohan. 2022. Long context question answering via supervised contrastive learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2872–2879, Seattle, United States. Association for Computational Linguistics.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8823–8838, Online. Association for Computational Linguistics.

Yifan Gao, Henghui Zhu, Patrick Ng, Cicero Nogueira dos Santos, Zhiguo Wang, Feng Nan, Dejiao Zhang, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Answering ambiguous questions through generative evidence fusion and round-trip prediction. 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 3263–3276, Online. Association for Computational Linguistics.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.

Vishal Gupta, Manoj Chinnakotla, and Manish Shrivastava. 2018. Retrieve and re-rank: A simple and effective IR approach to simple question answering over knowledge graphs. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 22–27, Brussels, Belgium. Association for Computational Linguistics.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Minda Hu, Muzhi Li, Yasheng Wang, and Irwin King. 2022. Momentum contrastive pre-training for question answering. *arXiv preprint arXiv:2212.05762*.

Maor Ivgi, Uri Shaham, and Jonathan Berant. 2023. Efficient long-text understanding with short-text models. *Transactions of the Association for Computational Linguistics*, 11:284–299.

Srinivasan Iyer, Sewon Min, Yashar Mehdad, and Wentau Yih. 2021. RECONSIDER: Improved re-ranking using span-focused cross-attention for open domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1280–1287, Online. Association for Computational Linguistics.

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.

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.

Bernhard Kratzwald, Anna Eigenmann, and Stefan Feuerriegel. 2019. Rankqa: Neural question answering with answer re-ranking. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. 2020. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3285–3292, 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. *arXiv preprint arXiv:1907.11692*.

Sagnik Majumder, Chinmoy Samant, and Greg Durrett. 2021. Model agnostic answer reranking system for adversarial question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 50–57, Online. Association for Computational Linguistics.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Reader-guided passage reranking for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 344–350, Online. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4498–4507, Online. Association for Computational Linguistics.

Sakib Shahriar and Kadhim Hayawi. 2023. Let's have a chat! a conversation with chatgpt: Technology, applications, and limitations. *arXiv preprint arXiv:2302.13817*.

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, Vancouver, Canada. Association for Computational Linguistics.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. 2021. Dense contrastive learning for self-supervised visual pre-training. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. 2022. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Nan Yang, Furu Wei, Binxing Jiao, Daxing Jiang, and Linjun Yang. 2021. xMoCo: Cross momentum contrastive learning for open-domain question answering. 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 6120–6129, Online. Association for Computational Linguistics.

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Zhenrui Yue, Bernhard Kratzwald, and Stefan Feuerriegel. 2021. Contrastive domain adaptation for question answering using limited text corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9575–9593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

Appendix

A Experimental result examples

A.1 Supervised ContrastiveQA examples

Data fields	Example
Answer	Mark Masons' Hall
Our proposed model answer prediction	Mark Masons' Hall
Baseline model answer prediction	86 St James's Street
Context	Mark Masons' Hall in London is the headquarters of The Grand Lodge of Mark Master Masons of England and Wales, which also controls the Royal Ark Mariner degree. It is located in **86 St James's Street** in the central London district of St James's, opposite St James's Palace. While Freemasons' Hall is the headquarters of the United Grand Lodge of England and the Supreme Grand Chapter of Royal Arch Masons of England, **Mark Masons' Hall** is the home of several other important appendant orders of Freemasonry in England and Wales. St James's Palace is the most senior royal palace in the United Kingdom. Located in the City of Westminster, although no longer the principal residence of the monarch, it is the ceremonial meeting place of the Accession Council and the London residence of several members of the royal family.
Question	What building is opposite the ceremonial meeting place of the Accession Council in the United Kingdom?
Answer	Adelaide
Our proposed model answer prediction	Adelaide
Baseline model answer prediction	Adelaide, about 4 km east of the Adelaide city centre. The suburb is in the City of Norwood
Context	Frewville is a small suburb in the South Australian city of Adelaide. It is three xa0kilometres south-east of Adelaide's central business district (CBD). Norwood is a suburb of **Adelaide, about 4 km east of the Adelaide city centre. The suburb is in the City of Norwood** Payneham & St Peters, the oldest South Australian local government municipality, with a city population over 34,000. Walter Frank Giffen (20 September 1861 in Norwood – 28 June 1949 in Adelaide) was an Australian cricketer who played in 3 Tests between 1887 and 1892. He was the brother of the great all-rounder George Giffen. Glenunga is a small southern suburb of 2,539 people in the South Australian city of Adelaide. It is located five kilometres southeast of the Adelaide city centre. Glenunga is a small southern suburb of 2,539 people in the South Australian city of Adelaide. It is located five kilometres southeast of the Adelaide city centre.
Question	Walter Giffen is from a suburb of which South Australian city?
Answer	2013
Our proposed model answer prediction	2013
Baseline model answer prediction	2015
Context	Frozen Fever is a **2015** American computer-animated musical fantasy short film produced by Walt Disney Animation Studios and released by Walt Disney Pictures. It is a sequel to the **2013** feature film "Frozen", and tells the story of Anna's birthday party given by Elsa with the help of Kristoff, Sven, and Olaf. Chris Buck and Jennifer Lee again served as the directors with Kristen Bell, Idina Menzel, Jonathan Groff, and Josh Gad providing the lead voices. "Making Today a Perfect Day" is a song from the 2015 Walt Disney Animation Studios computer-animated short film "Frozen Fever", with music and lyrics by Kristen Anderson-Lopez and Robert Lopez and performed throughout most of the short. It was released as a single in the United States on March 12, 2015.
Question	What is the year of the event that occured first, Making Today a Perfect Day was produced, or Frozen was produced?

Table A1: Example of answer extraction of Baseline and ContrastiveQA.

In Table A.1, we classify into positive and negative samples using the information of the given correct answer and train them with contrastive loss to derive a result that is closer to the original correct answer when compared with the baseline.