Leader-Generator Net: Dividing Skill and Implicitness for Conquering FairytaleQA

Wei Peng†
pengwei@iie.ac.cn
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China

Wanshui Li†
wanshui.li.19@ucl.ac.uk
University College London
London, United Kingdom

Yue Hu*
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China

ABSTRACT

Machine reading comprehension requires systems to understand the given passage and answer questions. Previous methods mainly focus on the interaction between the question and passage. However, they ignore the deep exploration of cognitive elements behind questions, such as fine-grained reading skills (this paper focuses on narrative comprehension skills) and implicitness or explicitness of the question (whether the answer can be found in the passage). Grounded in prior literature on reading comprehension, the understanding of a question is a complex process where human beings need to understand the semantics of the question, use different reading skills for different questions, and then judge the implicitness of the question. To this end, a simple but effective Leader-Generator Network is proposed to explicitly separate and extract fine-grained reading skills and the implicitness or explicitness of the question. Specifically, the proposed skill leader accurately captures the semantic representation of fine-grained reading skills with contrastive learning. And the implicitness-aware pointer-generator adaptively extracts or generates the answer based on the implicitness or explicitness of the question. Furthermore, to validate the generalizability of the methodology, we annotate a new dataset named NarrativeQA 1.1. Experiments on the FairytaleQA and NarrativeQA 1.1 show that the proposed model achieves the state-of-the-art performance (about 5% gain on Rouge-L) on the question answering task. Our annotated data and code are available at https://github.com/pengwei-iie/Leader-Generator-Net.

CCS CONCEPTS

• **Computing methodologies** → *Discourse, dialogue and pragmatics*; **Natural language processing**.

KEYWORDS

machine reading comprehension, reading skills, implicit or explicit question, question answering

*Corresponding author, e-mail:huyue@iie.ac.cn. † Equal Contribution.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591710

ACM Reference Format:
Wei Peng†, Wanshui Li†, and Yue Hu. 2023. Leader-Generator Net: Dividing Skill and Implicitness for Conquering FairytaleQA. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3539618.3591710

1 INTRODUCTION

Machine Reading Comprehension (MRC) has made significant strides [8, 26–28, 34, 41, 43], which focuses on teaching systems to understand the given passage and answer questions [8]. Different from the retrieval question answering system based on pattern matching, MRC is more about the ability of language comprehension and reasoning [19]. It is a critical step to train MRC systems that assist users in a variety of scenarios [19, 35], including voice assistant, intelligent customer service, search engine, etc.

In recent years, some researches [17, 34, 42] aim to make interactions between the question and passage from shallow to deep. Seo et al. [34] first introduce a bidirectional attention flow for MRC. Liao et al. [17] propose to construct a heterogeneous structure to make a deeper interaction. Most of these methods rely on the interaction between the question and passage or refining the key evidence from the passage. However, little is known about exploring the elements behind the question (e.g., fine-grained reading skills (details are in Section 2.1) and implicitness or explicitness of the question).

Grounded in prior literature in reading comprehension [14, 36], the understanding of a question is a complex process [11] where human beings need to understand the semantics of the question [14], use different reading skills [36, 39] (e.g., action or character) for different questions, and then judge the implicitness of the question [41]. Specifically, Purves et al. [29] and Vähäpassi [37] divide the reading task into three levels, each of which examines different reading skills that are utilized to answer the corresponding questions in a targeted way. In addition, implicitness or explicitness of the question indicates whether the answer can be extracted directly from the given passage, which can determine how the model outputs the answer [41]. As shown in Figure 1, an example is presented to illustrate the above process. Given the question and passage, the system should explore the cognitive elements in the step one, then make an answer in the step two. In this example, the reading skill is *Action* which indicates that *the skill could be utilized to answer about behaviors or information about that behavior.* And the question is explicit, which means the answer could be extracted from the passage. Therefore, the understanding of the question should not only

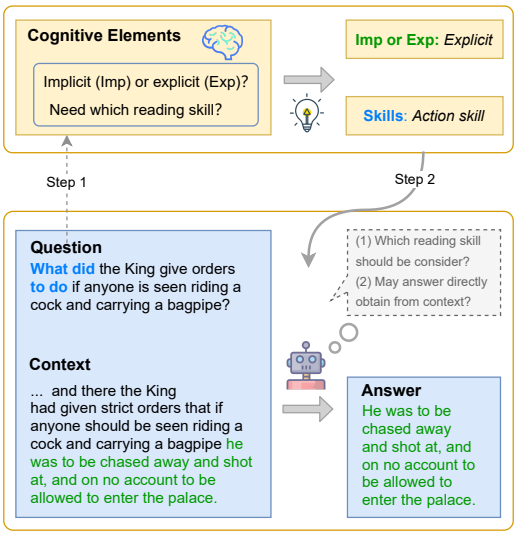


Figure 1: An example from FairytaleQA. In the first step, the system simulates the reading process of humans. Then, it predicts an answer in the second step.

be limited to its context, but also needs to excavate fine-grained reading skills and the implicitness or explicitness of the question.

In this paper, we propose a novel Leader-Generator Network (LGNet) to separate and extract the fine-grained reading skills and the implicitness or explicitness of the question. The proposed model consists of a skill leader, context encoder, implicitness encoder and implicitness-aware pointer-generator. Specifically, the skill leader aims to capture the difference and semantic representation of fine-grained reading skills with contrastive learning. The context encoder based on the pre-trained language model is used to obtain contextual information. Then, the implicitness encoder learns to determine whether the question is implicit, which is leveraged in the implicitness-aware pointer-generator to output the final answer.

The contributions can be summarized as follows:

- Mimicking human reading processes, the proposed LGNet explicitly explores two elements from the question, including fine-grained reading skills and implicitness or explicitness.
- The Skill Leader based on contrastive learning can accurately capture the difference and semantics of fine-grained reading skills.
- The Implicitness-aware Pointer-Generator can adaptively extract or generate the answer based on the implicitness or explicitness of the question.
- We annotate a new dataset called NarrativeQA 1.1. Experiments on two datasets show that the proposed model achieves the state-of-the-art (SOTA) performance significantly.

2 RELATED WORK

2.1 Narrative MRC Datasets

Recently, MRC has attracted much interest with a variety of datasets, such as SQuAD [31], NarrativeQA [12] and FairytaleQA [41]. Nowadays, research into narrative question answering [12, 13, 41] has grown rapidly, whose narrative writing style of book stories differs from the formal texts in Wikipedia and news, which demands a deeper understanding capability [21]. Considering the rich annotation in the FairytaleQA [41] whose reading skills and implicitness or explicitness of the question can be utilized, we focus on the question answering task on FairytaleQA. To validate the generalizability of the methodology, we annotate a new dataset named NarrativeQA 1.1 which contains the same reading skills as FairytaleQA. Specifically, the definitions of each reading skill is described as follows.

Character questions ask test takers to identify the character of the story or describe characteristics of characters.

Setting questions ask about a place or time where/when story events take place and typically start with "*Where*" or "*When*".

Action questions ask about characters' behaviors or information about that behavior.

Feeling questions ask about the character's emotional status or reaction to certain events and are typically worded as "*How did/does/do ...feel*".

Causal relationship questions focus on two events that are causally related where the prior events causally lead to the latter event in the question. This type of questions usually begins with "*Why*" or "*What made/makes*".

Outcome resolution questions ask for identifying outcome events that are causally led to by the prior event in the question. This type of questions are usually worded as "*What happened/happens/has happened...after...*".

Prediction questions ask for the unknown outcome of a focal event, which is predictable based on the existing information in the text.

2.2 MRC Models

In general, MRC models contain four key modules [19]: embeddings, feature extraction, context-question interaction and answer prediction. And main-stream studies mainly focus on the embedding layer [44, 45] or context-question interaction layer [17, 34]. For instance, Zhang et al. [45] introduce the syntax-guided information on MRC. MHPGM [2] uses a multi-attention mechanism to perform multiple hops of reasoning on narrative qa. Liao et al. [17] propose to leverage the heterogeneous structure to make a deeper interaction. These methods focus on using the external knowledge base or capturing the key evidence from the passage. However, little is known about exploring the cognitive elements behind the question. In this paper, we focus on explicitly extracting fine-grained reading skills and implicitness or explicitness of the question.

2.3 Contrastive Learning

Contrastive learning can map positive pairs closer, while pushing apart negative pairs, which has been widely utilized in NLP tasks [3, 6, 7, 10]. SimCSE [6] designs a simple approach that predicts the input sentence itself. DeCLUTR [7] presents a self-supervised objective for learning universal sentence embeddings. In addition, some researches [9, 16] propose a loss for supervised learning by leveraging label information. PairSCL [16] designs a supervised contrastive learning model to learn the sentence embedding. Our method is structurally similar to that used in [16] for supervised contrastive learning, with modifications for supervised classification.

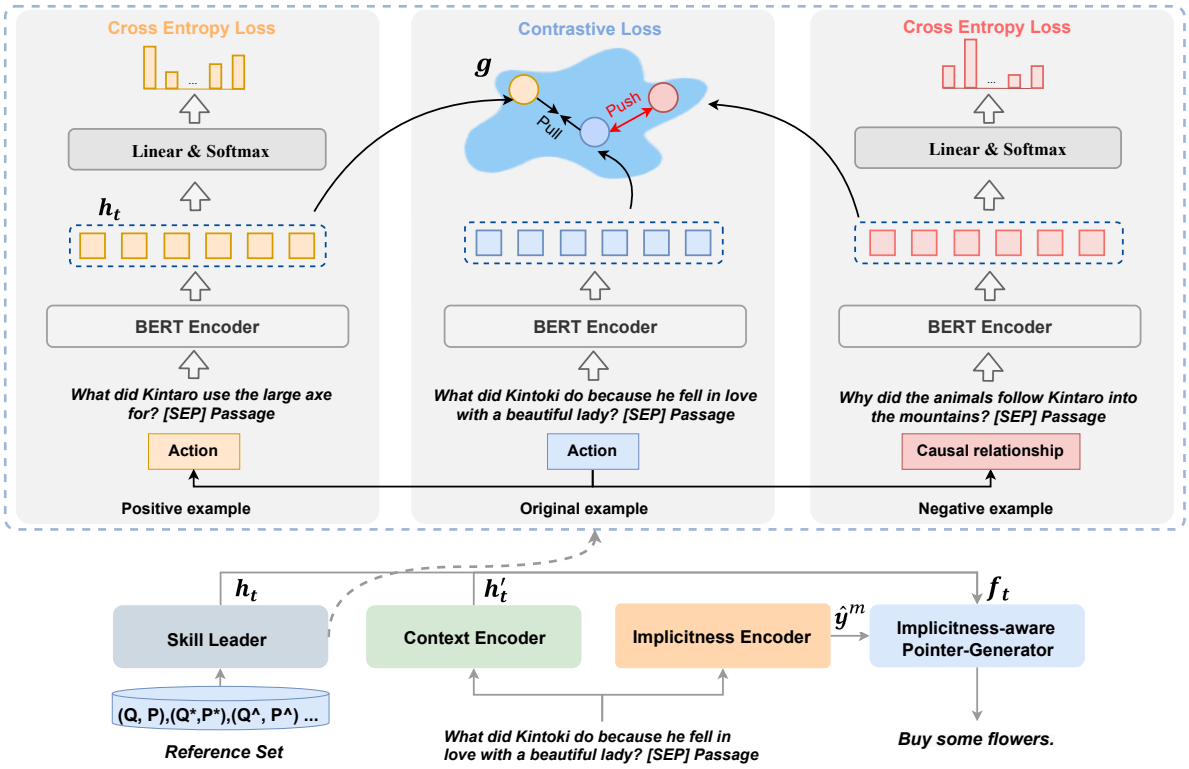


Figure 2: The overview of our framework. The dotted box represents the details of the skill leader.

3 PROBLEM FORMULATION

Given a question $X^q = \{x_1^q, x_2^q, \ldots, x_N^q\}$ and a passage $X^p = \{x_1^p, x_2^p, \ldots, x_M^p\}$, where N and M mean the length of question and passage, respectively. The answer has O words $X^a = \{x_1^a, x_2^a, \ldots, x_O^a\}$, and corresponding skill label and implicitness label of the question are y^s and y^m, respectively. The proposed model generates an output sequence $Y = \{y_1, y_2, \ldots, y_Z\}$ conditioned on the separated fine-grained reading skills and the implicitness or explicitness of the question, where z is the length of the answer.

4 APPROACH

The framework is described in Figure 2, which consists of the Skill Leader, Context Encoder, Implicitness Encoder and Implicitness-aware Pointer-Generator. The training of the model is divided into two stages. In the first stage, the skill leader makes a pretraining to learn the semantics of fine-grained reading skills with contrastive learning, which will be leveraged to make a guidance in stage two. In the second stage, the context encoder obtains the contextual representation between the question and passage. Then, the implicitness encoder aims to distinguish implicitness from explicitness of the question. Finally, implicitness-aware pointer-generator outputs the answer with the separated reading skills and implicitness or explicitness of the question.

4.1 Training Stage I

In the first training stage, we make a pretraining to the skill leader.

4.1.1 Skill Leader. Based on supervised contrastive learning, the proposed skill leader can accurately capture the difference and semantics of fine-grained reading skills that are utilized to make a guidance to answer the different questions with different skills. The details are as follows.

Skill References. Considering the rich annotation of the question in the FairytaleQA, the straightforward approach to generate positive samples is to annotate questions that explicitly require the same reading skills. Therefore, we directly construct the positive reference set with the same labels. The negative reference set can be randomly sampled with different class.

Skill Supervised Contrastive Learning. Motivated by the researches that extend the self-supervised batch contrastive approach to the fully-supervised setting to effectively leverage label information, we adopt supervised contrastive learning objective to accurately capture the semantic representation of the reading skills.

Given the reference set that contains batch-size instances $\mathcal{B}$, each of them is denoted as $(X^q, X^p, y^s)_{i \in \mathcal{B}}$, where $i = \{1, \ldots, K\}$ is the indices of the samples and K is the batch-size, y^s indicates the reading skill label of the instance. Following BERT [5], we add the start token ([CLS]) and separate the question X^q and passage X^p with a special token ([SEP]), the total length of the input is $T = (N + M + 2)$. The BERT encoder takes X^q, X^p as inputs and computes the contextual representations, leading to a series of contextual hidden states $(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T)$, as:

$$\boldsymbol{h}_t = \text{BERT}([\text{CLS}], X^q, [\text{SEP}], X^p) \quad (1)$$

where T is the total length of the input, $\boldsymbol{h}_t \in R^d$ is the t-th token in the input, and d is the hidden size.

The summary representation of the input, defined as $\boldsymbol{g}$, can be computed via a mean-pooling operation on all token representations:

$$\boldsymbol{g} = \texttt{Mean-pooling}(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T) \tag{2}$$

In the first training stage, we randomly sample a batch $\mathcal{I}$ of K examples $(X^q, X^p, y^s)_{i \in \mathcal{I}=\{1,\ldots,K\}}$ as described above. We denote the set of positives as $\mathcal{P} = \{p : p \in \mathcal{I}, y_p^s = y_i^s \wedge p \neq i\}$, with size $|\mathcal{P}|$. The supervised contrastive loss on the batch $\mathcal{I}$ is defined in the following:

$$\ell_{i,p} = \frac{\texttt{exp}(\texttt{sim}(\boldsymbol{g}_i, \boldsymbol{g}_p)/\tau)}{\sum_{k \in \mathcal{I}/i} \texttt{exp}(\texttt{sim}(\boldsymbol{g}_i, \boldsymbol{g}_k)/\tau)} \tag{3}$$

where $\boldsymbol{g}_i$ means the summary representation of the i-th sample. $\text{sim}(\boldsymbol{g}_i, \boldsymbol{g}_p) = (\boldsymbol{g}_i^T \cdot \boldsymbol{g}_p)$ denotes the cosine similarity of two vectors. $\ell_{i,p}$ indicates the likelihood that sample i is most relevant with the sample p, and τ is the temperature hyper-parameter. The larger the τ value, the smaller the dot product, and the more difficult the comparison becomes.

After obtaining $\ell_{i,p}$, the supervised contrastive loss $\mathcal{L}_{SCL}$ is obtained for every sample among the batch $\mathcal{I}$, which can be calculated as follows:

$$\mathcal{L}_{SCL} = \sum_{i \in \mathcal{I}} -\texttt{log} \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \ell_{i,p} \tag{4}$$

Cross-entropy Loss. The cross-entropy loss is also leveraged to consider the explicit force in discriminating various classes, as:

$$\mathcal{L}_{CE} = \texttt{CrossEntropy}\,(\texttt{MLP}(\boldsymbol{g}), y^s) \tag{5}$$

Finally, the overall loss of first training stage is an average of CE and the SCL loss, as:

$$\mathcal{L}_s = \mathcal{L}_{SCL} + \mathcal{L}_{CE} \tag{6}$$

4.2 Training Stage II

In the second training stage, the parameter of the skill leader is fixed. We focus on the question answering.

4.2.1 Context Encoder. The context encoder $\texttt{Enc}_{\texttt{ctx}}$ captures the contextual representations of the question and passage. Similarly, $\texttt{Enc}_{\texttt{ctx}}$ is utilized to encode each token to obtain contextual representations $(\boldsymbol{h'}_1, \ldots, \boldsymbol{h'}_T)$, like Equation 1. Then, to enable the model to adaptively answer different questions based on different reading skills, we incorporate the skill-aware representation $\boldsymbol{h}_t$ (in Equation 1) to make a guidance for subsequent prediction, as:

$$\boldsymbol{f_t} = \texttt{MLP}\,(\boldsymbol{h}_t, \boldsymbol{h}'_t) \tag{7}$$

where $\boldsymbol{f_t}$ is the representation which combines the contextual information and semantics of reading skills.

4.2.2 Implicitness Encoder. The implicitness encoder focuses on capturing the implicit or explicit relationship between the question and passage. Generally, explicit questions revolve around a specific fact in the text, and implicit questions require the evidence that is only implicit behind the context.

Specifically, the representation of implicit relationship between the question and passage $\tilde{\boldsymbol{h}}_t$ is calculated by $\texttt{Enc}_{\texttt{imp}}$ that has the same architecture but differnet parameters with $\texttt{Enc}_{\texttt{ctx}}$, as:

$$\tilde{\boldsymbol{h}}_t = \texttt{Enc}_{\texttt{imp}}([\texttt{CLS}], X^q, [\texttt{SEP}], X^p) \tag{8}$$

Then, a softmax function predicts the probability distribution of implicitness of the question $\hat{\boldsymbol{y}}^m$, as:

$$\hat{\boldsymbol{y}}^m = \texttt{Softmax}(\boldsymbol{W}_m^T \tilde{\boldsymbol{h}}_1 + \boldsymbol{b}_m) \tag{9}$$

where $\boldsymbol{W}_m \in R^{d \times j}$, $\boldsymbol{b}_m \in R^j$, $\hat{\boldsymbol{y}}^m \in R^j$, $j = 2$, which means the probability distribution that the question belongs to implicit or explicit, $\tilde{\boldsymbol{h}}_1$ is the [CLS] representation.

4.2.3 Implicitness-aware Pointer-Generator. The implicitness-aware pointer-generator is able to adaptively copy words from the source text, while simultaneously retaining the ability to generate novel words through the vocabulary. Inspired by the work PGNet [33], the implicitness-aware pointer-generator makes an advancement in the calculation on the generation probability $p_{\texttt{gen}} \in [0, 1]$. The original work leverages the context vector to obtain the $p_{\texttt{gen}}$ which is used as a soft switch to choose between *generating* a word from the vocabulary, or *copying* a word from the input sequence. However, this method only relies on hidden states of context, and it is also lacking explainability. Intuitively, implicitness or explicitness of the question indicates whether the answer can be extracted directly from the given passage or generated, which can determine how the model outputs the answer. Therefore, the proposed implicitness-aware pointer-generator regards the probability distribution $\hat{\boldsymbol{y}}^m \in R^j$ $(j = 2)$ (in Euqation 9) as the soft switch $p_{\texttt{gen}}$ and $(1 - p_{\texttt{gen}})$.

Specifically, the probability of the implicitness can be defined as $p_{\texttt{gen}}$, and the probability of the explicitness is regarded as $(1 - p_{\texttt{gen}})$. We obtain the final probability distribution over the extended vocabulary in the following:

$$\boldsymbol{p}\left(y_z | \boldsymbol{y}_{<z}, \boldsymbol{f}_t, \hat{\boldsymbol{y}}^m\right) = p_{\texttt{gen}} P_{\texttt{vocab}} + (1 - p_{\texttt{gen}}) \boldsymbol{\alpha}_z \tag{10}$$

where z indicates the decoding timestep, $\boldsymbol{f}_t$ can be seen in Equation 7, $\hat{\boldsymbol{y}}^m$ can be seen in Euqation 9. $\boldsymbol{\alpha}_z$ denotes the cross attention of last hidden layer in the context encoder in the z-th decoding timestep. And $P_{\texttt{vocab}}$ is a probability distribution over all words in the vocabulary, as:

$$P_{\texttt{vocab}} = \texttt{Decoder}(\boldsymbol{W}_{y<z}, \boldsymbol{f}_t) \tag{11}$$

where $\boldsymbol{W}_{y<z}$ denotes the embeddings of the generated tokens.

The cross-entropy loss of the implicitness recognition and answer generation are optimized as:

$$\mathcal{L}_1 = -y^m \log \hat{\boldsymbol{y}}^m \tag{12}$$

$$\mathcal{L}_2 = -\sum_{z=1}^{Z} \log \boldsymbol{p}\left(y_z | \boldsymbol{y}_{<z}, \boldsymbol{f}_t, \hat{\boldsymbol{y}}^m\right) \tag{13}$$

where y^m is the implicitness label of the question.

In the second training stage, we combine the above two loss functions as the final joint objective, as: $\mathcal{L}_s = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2$. λ_1 and λ_2 are two hyper-parameters for controlling the weight of the tasks.

Table 1: Core statistics of the FairytaleQA, which has 278 books and 10,580 QA-pairs. S.D. presents Standard Deviation.

FairytaleQA Dataset	Total			
	278 Books with 10,580 QA-pairs			
	Mean	S.D.	Min	Max
# section per story	15.6	9.8	2	60
# tokens per story	2305.4	1480.8	228	7577
# tokens per section	147.7	60.0	12	447
# questions per story	41.7	29.1	5	161
# questions per section	2.9	2.380	0	18
# tokens per question	10.5	3.2	3	27
# tokens per answer	7.2	5.8	1	70

5 EXPERIMENT

5.1 Dataset

To demonstrate the effectiveness of our work, FairytaleQA benchmark dataset is mainly chosen. Furthermore, to validate the generalizability of the proposed model, another narrative-related dataset, NarrativeQA, is annotated with the same reading skills as FairytaleQA, which is called NarrativeQA 1.1.

NarrativeQA 1.1. NarrativeQA [12] is one of the narrative-related datasets, which is generated by crowd workers who wrote QA pairs according to summaries of books or movie scripts. However, the original dataset lacks of the reading skill label. To further validate the generalizability of Leader-Generator Net, we employ three annotators who have background knowledge to label the dataset, and then construct a new dataset called NarrativeQA 1.1. Considering the questions are all implicit, annotators ignore this factor. The final results are determined by majority voting, the data is retained when the results of two of them are consistent. In case three annotators reached three different conclusions, the example will be discarded. Finally, NarrativeQA 1.1 consists of 21,841 questions (9,490 training examples, 3,461 validation examples and 8,890 test examples), covering seven types of narrative reading skills. And the constructed dataset will be released after reviewing for further research [1].

FairytaleQA. We evaluate our model and compared approaches on the FairytaleQA dataset [41], a dataset focusing on narrative comprehension generated by educational experts. The dataset consists of 10,580 explicit and implicit questions, covering seven types of narrative reading skills (details can be seen in Section 2.1). Following the official dataset [41], the train/validation/test dataset partition is split with a QA ratio of roughly 8:1:1. The overall statistics of the FairytaleQA examples are shown in Table 1.

5.2 Evaluation Metrics

(1) For the answer generation task, the BLEU-n (B-n) [24], Rouge-L (R-L) [18] and METEOR [1] are utilized as our main metrics, which are widely used for evaluating the quality of language generation. Following the paper [41], we use background color to highlight the column of Rouge-L results in Table 2. (2) For the skill recognition task and implicitness determination task, prediction accuracy (ACC) is leveraged in the analysis experiment.

[1] https://github.com/pengwei-iie/Leader-Generator-Net

5.3 Experimental Setting

We utilize the BART [15] (BART-large) as the context encoder and decoder following paper [41] with Pytorch [25] framework. Since BART outperforms other model architectures in the QA task of FairytaleQA [41], we decide to use BART-large as the backbone for our fine-tuned Leader-Generator Network. For FairytaleQA, the epoch is set to 10 with the learning rate as 5e-6 and a linear warmup with 100 steps. The batch size of training is 1. The AdamW [20] optimizer is used for training with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e$-8. The hyper-parameters λ_1 and λ_2 in the training objective are set to 0.5 and 0.5, respectively. As for the first training stage, the BERT-base is utilized. The batch size is set to 32, with the learning rate $5e-5$. Other parameters are the same as the paper [16].

For NarrativeQA 1.1, considering the questions are all implicit, the hyper-parameters λ_1 and λ_2 are set to 0 and 1, respectively. The epoch is set to 3 and the learning rate is set to 5e-5. All other remaining parameters are the same as the setting of FairytaleQA. We use Tesla V100-32G GPU to implement our experiments. The source code[2] is released to facilitate future work.

5.4 Baselines

We provide baselines that contain three types, including lightweight models (trained with FairytaleQA), Pre-trained Language Models (PLMs) (directly evaluate, following the paper [41]) and fine-tuned models for comparison.

Seq2Seq [4]: A RNN Encoder–Decoder framework that widely used in the natural language generation task.

Transformer [38]: The standard Transformer model.

BERT [5]: A widely used pre-trained language model on comprehension tasks.

DistilBERT [32]: A variation of BERT which is a smaller PLM.

Masque [22]: Masque is based on multi-source abstractive summarization and learns multi-style answers together.

BART [15]: A sequence-to-sequence framework which is effective on text generation and comprehension tasks.

T5 [30]: T5 is a unified textto-text pre-training framework that covers all text-based language problems. We use T5-base from HuggingFace Transformers [40].

ProQA [46]: ProQA is a unified QA paradigm that takes a unified structural prompt as the bridge and improves the QA-centric ability.

Pea-QA [23]: Pea-QA studies parameter-efficient transfer learning over text for abstractive question answering using adapters.

BART on NarrativeQA [41]: A variation of the BART fine-tuned on NarrativeQA dataset, which is a strong baseline. Note: we have done two variations during the experiment in NarrativeQA 1.1. One is the model that fine-tuned on NarrativeQA. Another is the approach which is reproduced by adding the skill recognition loss for a fair comparison (as shown in Table 2 *) with ours.

BART on FairytaleQA [41]: BART fine-tuned on FairytaleQA dataset, which is the SOTA model. Note: we also have done two variations. One is the model in the paper [41] (as shown in Table 2 †). Another is the approach which is reproduced by adding the skill recognition loss and implicitness determination loss for a fair comparison (as shown in Table 2 ‡) with ours (the number of parameters is almost the same as the LGNet).

Table 2: Performance on the validation and the test set on FairytaleQA and NarrativeQA 1.1. We report the results of the BLEU-*n*, Rouge-L and METERO. † refers to the method in the paper [41], which is the SOTA model. ‡ refers to the approach which is reproduced with the skill recognition loss and implicitness determination loss on FairytaleQA for a fair comparison with ours. ∗ refers to the approach which is reproduced with the skill recognition loss on NarrativeQA 1.1. We use background color to highlight the column of Rouge-L results [41].

	NarrativeQA 1.1 Validation						NarrativeQA 1.1 Test					
	B-1	**B-2**	**B-3**	**B-4**	**Rouge-L**	**METEOR**	**B-1**	**B-2**	**B-3**	**B-4**	**Rouge-L**	**METEOR**
Lightweight Models												
Seq2Seq [4]	16.10	-	-	1.40	13.29	4.22	15.89	-	-	1.26	13.15	4.08
Transformer [38]	23.10	6.01	2.64	1.83	11.51	6.47	21.08	4.87	1.98	0.85	10.94	5.89
Fine-tuned Models												
ProQA [46]	-	-	-	-	-	-	-	-	-	-	50.10	-
Pea-QA [23]	-	-	-	-	-	-	-	-	-	-	51.50	-
Masque [22]	-	-	-	-	-	-	-	-	-	-	54.70	-
T5 [30]	47.58	33.72	21.66	13.25	48.27	21.56	49.25	35.91	23.18	16.30	49.43	22.43
BART-NarrativeQA [41]	51.52	38.68	29.28	22.26	55.92	25.92	53.66	41.57	32.57	25.76	56.79	26.51
BART-NarrativeQA [41] *	55.50	41.31	31.92	25.08	56.84	26.36	54.04	43.59	34.28	27.04	57.09	27.35
Leader-Generator Net (LGNet)	**56.29**	**43.67**	**34.03**	**26.66**	**61.16**	**28.51**	**58.12**	**46.46**	**37.35**	**30.16**	**62.20**	**28.83**

	FairytaleQA Validation						FairytaleQA Test					
	B-1	**B-2**	**B-3**	**B-4**	**Rouge-L**	**METEOR**	**B-1**	**B-2**	**B-3**	**B-4**	**Rouge-L**	**METEOR**
Lightweight Models												
Seq2Seq [4]	25.12	6.67	2.01	0.81	13.61	6.94	26.33	6.72	2.17	0.90	14.55	7.34
Transformer [38]	21.87	4.94	1.53	0.59	10.32	6.01	21.72	5.21	1.74	0.67	10.27	6.22
Pre-trained Language Models												
DistilBERT [32]	-	-	-	-	9.70	-	-	-	-	-	8.20	-
BERT [5]	-	-	-	-	10.40	-	-	-	-	-	9.70	-
BART [15]	19.13	7.92	3.42	2.14	12.25	6.51	21.05	8.93	3.90	2.52	12.66	6.70
Fine-tuned Models												
T5 [30]	41.39	35.22	33.16	31.05	43.19	22.17	43.28	37.07	34.56	32.10	43.76	23.58
BART-NarrativeQA [41]	45.34	39.17	36.33	34.10	47.39	24.65	48.13	41.50	38.26	36.97	49.16	26.93
BART-FairytaleQA [41] †	50.55	43.30	41.23	38.29	53.88	27.09	54.04	45.98	42.08	39.46	53.64	27.45
BART-FairytaleQA [41] ‡	51.57	44.49	41.96	39.67	54.76	27.14	55.21	47.08	43.57	40.34	55.61	28.11
Leader-Generator Net (LGNet)	**55.78**	**49.70**	**46.66**	**44.48**	**59.83**	**30.32**	**58.43**	**51.92**	**48.54**	**45.96**	**60.97**	**31.36**
Human [41]	-	-	-	-	65.10	-	-	-	-	-	64.40	-

5.5 Main Comparison

Results on NarrativeQA 1.1. For NarrativeQA 1.1 dataset, SOTA models [22, 23, 46] and PLMs [30, 41] are as baselines to compare with our LGNet. From Table 2, it can be concluded that the performance of our LGNet outperforms baselines remarkably. In addition, to make a fair comparison, BART-NarrativeQA* is reproduced by adding skill recognition loss with a softmax function. Apparently, compared with BART-NarrativeQA, our model achieves about 5.41% gain on Rouge-L on the test set, which suggests that capturing the semantics of reading skills is beneficial for the QA task.

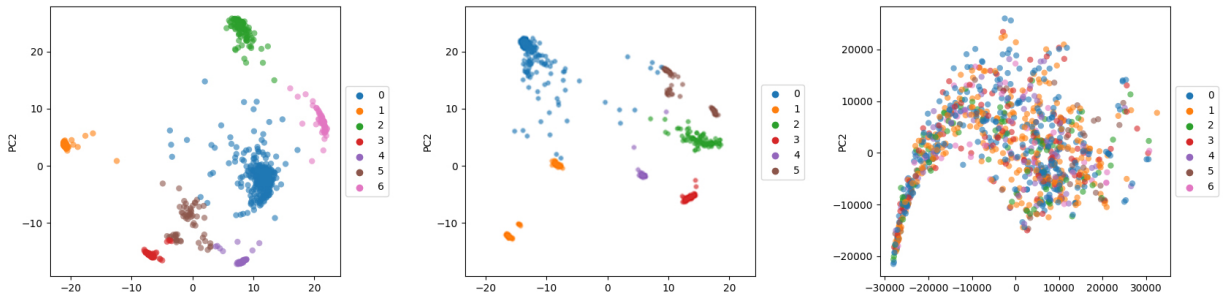
Results on FairytaleQA. The main results on the validation set and the test set of the FairytaleQA dataset are shown in Table 2. Note that, the evaluation metric is only Rouge-L in the paper [41]. Therefore, the results are reproduced by ourselves in the experiments, and we achieve consistent performance with the paper. As shown in Table 2, for lightweight models (Seq2Seq and Transformer) and PLMs (directly evaluate, following the paper [41]), the performance is relatively low, which indicates their insufficient comprehension ability on narrative reading skills, and they are still far from human performance. Compared to SOTA models, LGNet obtains SOTA performance on all the evaluation metrics significantly, which achieves 5.62% gain on B-4 and 5.36% gain on Rouge-L on the test set. It is worth noting that the LGNet has closed the gap to 4% compared with human performance, which is smaller than the report (around 12%) in the paper [41].

6 ANALYSES

In this section, to validate the effectiveness of the model's components, we conduct an ablation study on the two datasets. In other analyses, we focus on the new and challenging dataset FairytaleQA. Some analyses on NarrativeQA 1.1 can be seen in the Appendix.

Table 3: Results of ablation study on each component on the validation and test set of the FairytaleQA and NarrativeQA 1.1.

	NarrativeQA 1.1 Validation						NarrativeQA 1.1 Test					
	B-1	**B-2**	**B-3**	**B-4**	**Rouge-L**	**METEOR**	**B-1**	**B-2**	**B-3**	**B-4**	**Rouge-L**	**METEOR**
BART-NarrativeQA	51.52	38.68	29.28	22.26	55.92	25.92	53.66	41.57	32.57	25.76	56.79	26.51
w/o Skl Led	51.52	38.68	29.28	22.26	55.92	25.92	53.66	41.57	32.57	25.76	56.79	26.51
w/o CL	55.50	41.31	31.92	25.08	56.84	26.36	54.04	43.59	34.28	27.04	57.09	27.35
Ours	**56.29**	**43.67**	**34.03**	**26.66**	**61.16**	**28.51**	**58.12**	**46.46**	**37.35**	**30.16**	**62.20**	**28.83**
	FairytaleQA Validation						**FairytaleQA Test**					
	B-1	**B-2**	**B-3**	**B-4**	**Rouge-L**	**METEOR**	**B-1**	**B-2**	**B-3**	**B-4**	**Rouge-L**	**METEOR**
BART-FairytaleQA‡	51.57	44.49	41.96	39.67	54.76	27.14	55.21	47.08	43.57	40.34	55.61	28.11
w/o Skl Led	52.01	44.93	42.30	40.55	56.02	28.77	54.47	47.32	43.73	41.10	57.72	30.29
w/o CL	53.52	46.02	43.11	41.08	57.41	29.03	55.75	48.37	44.50	41.62	58.45	30.67
w/o Imp Enc	53.48	46.23	43.39	42.36	58.21	28.15	56.14	48.77	45.07	41.67	57.38	29.70
Ours	**55.78**	**49.70**	**46.66**	**44.48**	**59.83**	**30.32**	**58.43**	**51.92**	**48.54**	**45.96**	**60.97**	**31.36**



(a) embeddings from LGNet. **(b) from BART-FairytaleQA ‡.** **(c) from the raw data space.**

Figure 3: The visualization of the skill recognition on FairytaleQA. PCA projections of the embeddings learned by different models. From left to right: projections learned (g as shown in Equation 2) using contrastive loss; projections learned with cross-entropy loss and without contrastive loss; raw data space (the original data without any embeddings).

6.1 Ablation Study

To get better insight into the components in LGNet, the ablation study is performed in Table 3. The experiments have shown that each component is beneficial to final results.

w/o Skill Leader (Skl Led) In this setting, the skill leader aims to accurately capture the semantic representation of fine-grained reading skills with the supervised contrastive learning. To evaluate the effectiveness of Skl Led, the skill leader is removed (i.e., without the contrastive loss and cross-entropy loss, as illustrated in Equation 4 and 5), leading to a significant decrease on all the metrics on these two datasets, which confirms that perceiving different reading skills for different questions is of great necessity. Note: removing this setting, the model degrades to BART-NarrativeQA model on NarrativeQA 1.1.

w/o Contrastive Learning (CL) The purpose of contrastive learning is to bring the semantic representations of samples belonging to the same class closer and push apart unrelated samples. And cross-entropy loss builds upon entropy and generally calculates the difference between two probability distributions. To verify that the decent performance of the model is not just derived from label information (cross-entropy loss), in this setting, we directly remove the contrastive loss (i.e., without Equation 4) and keep the cross-entropy loss. As shown in Table 3 line 5 and line 11, the performance has declined on these two datasets, which proves that contrastive learning has the potential to make a more precise recognition of different reading skills.

w/o Implicitness Encoder (Imp Enc) The implicitness encoder focuses on capturing the implicit relationship between the question and passage, so as to determine how the model outputs the answer by the implicitness-aware pointer-generator. *Note: the implicitness-aware pointer-generator will automatically degrade to the BART decoder in this setting.* As shown in Table 3 (line 12), removing this factor leads to a deeper impact on the FairytaleQA dataset, which demonstrates that the module makes a contribution to the overall improvement. Note: there is no Imp Enc setting in NarrativeQA 1.1 because the questions in NarrativeQA 1.1 are all implicit.

Table 4: Performance of the skill recognition and question answering task on cross-entropy-based model and contrastive learning on FairytaleQA test set.

	ACC ↑	B-4 ↑	R-L ↑	METEOR ↑
Cross-entropy-based	92.28	41.62	58.45	30.67
Contrastive Learning	**94.13**	**45.96**	**60.97**	**31.36**

Table 5: Analysis comparison of different decoders on FairytaleQA test set.

	ACC ↑	B-4 ↑	R-L ↑	METEOR ↑
Original BART Decoder	-	41.67	57.38	29.70
Original PGNet	-	41.73	58.62	29.16
LGNet	**75.65**	**45.96**	**60.97**	**31.36**

6.2 Skill Visualization Analysis

To confirm the effectiveness of the supervised contrastive learning, we check the PCA projections of the embeddings ($\boldsymbol{g}$ as shown in Equation 2) from the proposed model with the contrastive loss, and from that learned using cross-entropy loss (the BART on FairytaleQA ‡). As a negative control, we also give the PCA projection of the original data space without any embeddings. As the scatter plots show, both the contrastive model and cross-entropy-based model cluster the samples with the same labels better than the original data, but the clusters in the contrastive embedding are much more distinguishable and the different categories are further separated in the semantic space. Furthermore, an interesting phenomenon is that the cross-entropy-based model only performs six clusters of samples rather than the true seven clusters.

In summary, the supervised contrastive learning can obtain better embedding space, thus obtain the better improvements on skill recognition and question answering tasks.

6.3 Skill Recognition Analysis

To explore how the contrastive learning affects the performance of skill recognition, and thus on question answering task, we present an analysis compared to the cross-entropy-based model (remove the $\mathcal{L}_{SCL}$ in Equation 6) in Table 4. As can be seen, our proposed model has obtained promising performance on both of the tasks. Specifically, LGNet achieves a better result which gains 1.85% on the ACC metric and 4.34% on the B-4 metric. And the higher accuracy of the skill prediction suggests that the contrastive learning is beneficial to making an appropriate recognition, which also helps the question answering task. (Note: we also make an analysis to the input of the skill leader simply with the question rather than question and passage, which can be seen in Section 6.6.)

6.4 Generator Analysis

The implicitness-aware pointer-generator is an essential module to determine how the model outputs the answer. Therefore, we make an analysis to verify the performance of this module, the original BART decoder and original PGNet. As shown in Table 5, the accuracy of the implicitness prediction can be up to 75.65%, so as to make a guidance for the output way. The results in the original BART decoder and original PGNet are approaching, which indicates that leveraging the implicitness of the question as the soft gate in PGNet is important. In addition, to confirm the performance improvement of LGNet is derived from the way of implicitness modeling rather than the working mechanism of the PGNet, we compare the original PGNet with ours. The LGNet obtains better results on all the metrics than PGNet, which demonstrates the necessity of the implicitness modeling in our approach.

Table 6: Qualitative analyses of our model and baselines.

Input passage There was once an old man and his wife, who lived in a dear little cottage by the side of a burn. they were a very canty and contented couple, for they had enough to ... **Question** Who were a very canty and contented couple?
Predicted Skill: Character. ✓ **Predicted Implicitness or explicitness:** Explicitness. ✓
Ground-truth An old man and his wife.
Outputs ***Ours:*** An old man and his wife. ***BART-FairytaleQA‡:*** They had enough to live on, and plenty of things to do. ***Transformer:*** An old man. ***Seq2Seq:*** The old mouse.
Input passage The novel begins as the narrator, ..., nursery where her dead first son would have been raised. Occupied with the domestic management of the Baldry estate just outside London, the two are almost completely removed from ... **Question** How do Kitty and Chanchala remain away from the effects of war?
Predicted Skill: Action. ✓ **Predicted Implicitness or explicitness:** Implicitness. ✓
Ground-truth: By managing Baldry Estate.
Outputs ***Ours:*** They manage the estate. ***BART-NarrativeQA*:*** They are domestic managers. ***BART:*** Domestic management. ***Transformer:*** They went to school together.

6.5 Case Study

Table 6 shows examples from ours and baselines qualitatively. In case one, our model predicts the *character* skill and *explicitness of the question*. Therefore, it generates the answer *an old man and his wife*, which describes the character attribution and is consistent with the predicted reading skill. Furthermore, the generated answer overlaps with the passage (as shown in cyan) in a high degree because of the explicitness. However, the baselines generate the wrong answer *they had enough to live on, and plenty of things to do* and incomplete answer *an old man*. In case two, red indicates the key evidence. Transformer predicts an irrelevant and wrong answer *they went to school together*. BART-style baselines output

Table 7: Performance of the skill recognition on different inputs of skill leader on FairytaleQA.

	ACC on validation set ↑	ACC on test set ↑
Only with question	87.71	83.71
With question & passage	**94.35**	**94.13**

a *Noun*, i.e., *domestic managers*, whose description is inconsistent with the reading skill *action*. Particularly, our model obtains a decent and generative answer *they manage the estate* with the help of the predicted reading skill *action* and implicitness of the question. The case study confirms that excavating the fine-grained reading skills and the implicitness or explicitness of the question is beneficial and necessary on question answering task. Another bad case can be seen in Appendix B.

6.6 Analysis to the Input of the Skill Leader

To explore the performance on different inputs of skill leader, we make an analysis in the following setting: 1) the input of the skill leader is simply a question, and 2) the input consists of the question and passage. As shown in Table 7, interestingly, the performance of skill recognition is better when given the question and passage. Through this experimental result, we argue that the learning of the reading skill not only depends on the question, but also on the relationship between the question and passage. From the human beings' perspective, reading skills can be easily predicted simply with the question (since humans have learned a great deal of knowledge). However, for machines without prior knowledge, questions are usually short, resulting in less information. By introducing the passage, the information has been augmented, making it easier for machines to learn the contextual knowledge, so as to make more accurate skill recognition. Furthermore, this setting guarantees that the input is consistent during both training stages, which can alleviate the gap caused by the different training stages.

7 CONCLUSION

This paper concentrates on the deep understanding of questions to mimic the reading process of human beings. We present a simple but effective Leader-Generator Network (LGNet) to explicitly separate and extract fine-grained reading skills and implicitness or explicitness of the question on FairytaleQA. Furthermore, to validate the generalizability of the LGNet, we annotate a new dataset named NarrativeQA 1.1 with the corresponding reading skills. Quantitative results show that LGNet has achieved the SOTA performance and improved the result of skill recognition. The qualitative analyses also demonstrate the importance of each component in LGNet. For future work, how to model the deeper implicitness relationship between the question and passage and automatically annotate the reading skill are still worth researching.

A SKILL RECOGNITION ANALYSIS

To further explore how the contrastive learning affects the performance of skill recognition on NarrativeQA 1.1, we make another analysis. The conclusion is similar to the demonstration in Section 6.3. As shown in Table 8, LGNet has made a promising achievement on both of the tasks. Specifically, LGNet achieves a better result which gains 2.60% on the ACC metric and 5.11% on the Rouge-L metric. And the higher accuracy of the skill prediction suggests that contrastive learning is beneficial to making an appropriate recognition, which also helps the question answering task.

Table 8: Performance of the skill recognition and question answering task on cross-entropy-based model and contrastive learning on NarrativeQA 1.1 test set.

	ACC ↑	B-4 ↑	R-L ↑	METEOR ↑
Cross-entropy-based	91.48	27.04	57.09	27.35
Contrastive Learning	**94.05**	**30.16**	**62.20**	**28.83**

B BAD CASE STUDY

Table 9 shows another bad example. The LGNet copies the words from the passage and outputs the correct answer. However, the predicted reading skill is *Precition*, which is inconsistent with the true label *Outcome resolution*. The possible reason is that these two reading skills are semantically similar, thus confusing our model. Nonetheless, the LGNet still obtains the correct answer with the help of the context information and explicitness of the question. By contrast, the generations of baselines are all inferior.

Table 9: More qualitative analysis of our model and baselines.

Input passage ... and he left the little man standing there and went further on into the forest. There he began to cut down a tree, but before long he made a false stroke with his axe, and cut his own arm so badly that he was obliged to go home and have it bound up. Then the second son went to forest, and his ...
Question What happened to the eldest son because he made a false stroke with his axe?

Predicted Skill: Prediction. ✗
True Skill: Outcome resolution.
Predicted Implicitness or explicitness: Explicitness. ✓

Ground-truth Cut his own arm badly and was obliged to go home and have it bound up.

Outputs

Ours: He cut his own arm so badly that he was obliged to go home and have it bound up.

BART-FairytaleQA‡: He cut his arm badly.

Transformer: She felt something in her hand, and when she looked down, she saw a large sandwich of bread.

Seq2Seq: The the heart into.

Acknowledgement

We thank all anonymous reviewers for their constructive comments. This work is supported by the National Natural Science Foundation of China (Grant No. 62006222) and (Grant No.U21B2009). This research is also supported by the Strategic Priority Research Program of Chinese Academy of Science, Grant No.XDC02030400.

REFERENCES

[1] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss (Eds.). Association for Computational Linguistics, 65–72. https://aclanthology.org/W05-0909/

[2] Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for Generative Multi-Hop Question Answering Tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 4220–4230. https://doi.org/10.18653/v1/d18-1454

[3] Feilong Chen, Xiuyi Chen, Shuang Xu, and Bo Xu. 2022. Improving Cross-Modal Understanding in Visual Dialog Via Contrastive Learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7937–7941.

[4] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1724–1734. https://doi.org/10.3115/v1/d14-1179

[5] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[6] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

[7] John M. Giorgi, Osvald Nitski, Bo Wang, and Gary D. Bader. 2021. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 879–895. https://doi.org/10.18653/v1/2021.acl-long.72

[8] Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). 1693–1701. https://proceedings.neurips.cc/paper/2015/hash/afdec7005cc9f14302cd0474fd0f3c96-Abstract.html

[9] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html

[10] Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-Guided Contrastive Learning for BERT Sentence Representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 2528–2540. https://doi.org/10.18653/v1/2021.acl-long.197

[11] Young-Suk Grace Kim. 2017. Why the simple view of reading is not simplistic: Unpacking component skills of reading using a direct and indirect effect model of reading (DIER). *Scientific Studies of Reading* 21, 4 (2017), 310–333.

[12] Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. *Trans. Assoc. Comput. Linguistics* 6 (2018), 317–328. https://doi.org/10.1162/tacl_a_00023

[13] Yash Kumar Lal, Nathanael Chambers, Raymond J. Mooney, and Niranjan Balasubramanian. 2021. TellMeWhy: A Dataset for Answering Why-Questions in Narratives. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021 (Findings of ACL, Vol. ACL/IJCNLP 2021)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 596–610. https://doi.org/10.18653/v1/2021.findings-acl.53

[14] Wendy G. Lehnert. 1977. Human and Computational Question Answering. *Cogn. Sci.* 1, 1 (1977), 47–73. https://doi.org/10.1207/s15516709cog0101_3

[15] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[16] Shuang Li, Xuming Hu, Li Lin, and Lijie Wen. 2022. Pair-Level Supervised Contrastive Learning for Natural Language Inference. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. IEEE, 8237–8241. https://doi.org/10.1109/ICASSP43922.2022.9746499

[17] Jinzhi Liao, Xiang Zhao, Xinyi Li, Jiuyang Tang, and Bin Ge. 2022. Contrastive heterogeneous graphs learning for multi-hop machine reading comprehension. *World Wide Web* 25, 3 (2022), 1469–1487. https://doi.org/10.1007/s11280-021-00980-6

[18] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[19] Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. 2019. Neural Machine Reading Comprehension: Methods and Trends. *CoRR* abs/1907.01118 (2019). arXiv:1907.01118 http://arxiv.org/abs/1907.01118

[20] Ilya Loshchilov and Frank Hutter. 2017. Fixing Weight Decay Regularization in Adam. *ArXiv* (2017).

[21] Xiangyang Mou, Chenghao Yang, Mo Yu, Bingsheng Yao, Xiaoxiao Guo, Saloni Potdar, and Hui Su. 2021. Narrative Question Answering with Cutting-Edge Open-Domain QA Techniques: A Comprehensive Study. *Trans. Assoc. Comput. Linguistics* 9 (2021), 1032–1046. https://doi.org/10.1162/tacl_a_00411

[22] Kyosuke Nishida, Itsumi Saito, Kosuke Nishida, Kazutoshi Shinoda, Atsushi Otsuka, Hisako Asano, and Junji Tomita. 2019. Multi-style Generative Reading Comprehension. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 2273–2284. https://doi.org/10.18653/v1/p19-1220

[23] Vaishali Pal, E. Kanoulas, and M. de Rijke. 2022. Parameter-Efficient Abstractive Question Answering over Tables or Text. In *Workshop on Document-grounded Dialogue and Conversational Question Answering*.

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 311–318. https://doi.org/10.3115/1073083.1073135

[25] Adam Paszke, S. Gross, Soumith Chintala, G. Chanan, E. Yang, Zachary Devito, Zeming Lin, Alban Desmaison, L. Antiga, and A. Lerer. 2017. Automatic differentiation in PyTorch.

[26] Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Jing Yu, Yajing Sun, and Xiangpeng Wei. 2020. Bi-directional cognitive thinking network for machine reading comprehension. *arXiv preprint arXiv:2010.10286* (2020).

[27] Wei Peng, Yue Hu, Jing Yu, Luxi Xing, and Yuqiang Xie. 2021. APER: AdaPtive Evidence-driven Reasoning Network for machine reading comprehension with unanswerable questions. *Knowl. Based Syst.* 229 (2021), 107364. https://doi.org/10.1016/j.knosys.2021.107364

[28] Aleksandr Perevalov, Andreas Both, Dennis Diefenbach, and Axel-Cyrille Ngonga Ngomo. 2022. Can Machine Translation be a Reasonable Alternative for Multilingual Question Answering Systems over Knowledge Graphs?. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini (Eds.). ACM, 977–986. https://doi.org/10.1145/3485447.3511940

[29] Alan C Purves, A Söter, Sauli Takala, and A Vähäpassi. 1984. Towards a domain-referenced system for classifying composition assignments. *Research in the Teaching of English* (1984), 385–416.

[30] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv* abs/1910.10683 (2019).

[31] 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, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, Jian Su, Xavier Carreras, and Kevin Duh (Eds.). The Association for Computational Linguistics, 2383–2392. https://doi.org/10.18653/v1/d16-1264

[32] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR* abs/1910.01108 (2019). arXiv:1910.01108 http://arxiv.org/abs/1910.01108

[33] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, 1073–1083. https://doi.org/10.18653/v1/P17-1099

[34] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional Attention Flow for Machine Comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=HJ0UKP9ge

[35] Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172* (2019).

[36] Saku Sugawara, Hikaru Yokono, and Akiko Aizawa. 2017. Prerequisite Skills for Reading Comprehension: Multi-Perspective Analysis of MCTest Datasets and Systems. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, Satinder P. Singh and Shaul Markovitch (Eds.). AAAI Press, 3089–3096. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14614

[37] Anneli Vähäpassi. 1981. On the specification of the domain of school writing. *AFinLAn vuosikirja* (1981), 85–107.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[39] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1502.05698

[40] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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, EMNLP 2020 - Demos, Online, November 16-20, 2020*, Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[41] Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic Questions and Where to Find Them: FairytaleQA - An Authentic Dataset for Narrative Comprehension. In *ACL*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 447–460. https://doi.org/10.18653/v1/2022.acl-long.34

[42] Ming Yan, Jiangnan Xia, Chen Wu, Bin Bi, Zhongzhou Zhao, Ji Zhang, Luo Si, Rui Wang, Wei Wang, and Haiqing Chen. 2019. A Deep Cascade Model for Multi-Document Reading Comprehension. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 7354–7361. https://doi.org/10.1609/aaai.v33i01.33017354

[43] Yuanmeng Yan, Rumei Li, Sirui Wang, Hongzhi Zhang, Zan Daoguang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Large-Scale Relation Learning for Question Answering over Knowledge Bases with Pre-trained Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 3653–3660. https://doi.org/10.18653/v1/2021.emnlp-main.296

[44] Bishan Yang and Tom M. Mitchell. 2017. Leveraging Knowledge Bases in LSTMs for Improving Machine Reading. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, 1436–1446. https://doi.org/10.18653/v1/P17-1132

[45] Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020. SG-Net: Syntax-Guided Machine Reading Comprehension. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 9636–9643. https://ojs.aaai.org/index.php/AAAI/article/view/6511

[46] Wanjun Zhong, Yifan Gao, Ning Ding, Yujia Qin, Zhiyuan Liu, Ming Zhou, Jiahai Wang, Jian Yin, and Nan Duan. 2022. ProQA: Structural Prompt-based Pre-training for Unified Question Answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (Eds.). Association for Computational Linguistics, 4230–4243. https://doi.org/10.18653/v1/2022.naacl-main.313