Evidence Sentence Extraction for Machine Reading Comprehension

Hai Wang[1*] **Dian Yu**[2] **Kai Sun**[3*] **Jianshu Chen**[2]
Dong Yu[2] **David McAllester**[1] **Dan Roth**[4]
[1]Toyota Technological Institute at Chicago, Chicago, IL, USA
[2]Tencent AI Lab, Bellevue, WA, USA [3]Cornell, Ithaca, NY, USA
[4]University of Pennsylvania, Philadelphia, PA, USA
{haiwang,mcallester}@ttic.edu, ks985@cornell.edu,
{yudian,jianshuchen,dyu}@tencent.com, danroth@seas.upenn.edu

Abstract

Remarkable success has been achieved in the last few years on some limited machine reading comprehension (MRC) tasks. However, it is still difficult to interpret the predictions of existing MRC models. In this paper, we focus on extracting evidence sentences that can explain or support the answers of multiple-choice MRC tasks, where the majority of answer options cannot be directly extracted from reference documents.

Due to the lack of ground truth evidence sentence labels in most cases, we apply distant supervision to generate imperfect labels and then use them to train an evidence sentence extractor. To denoise the noisy labels, we apply a recently proposed deep probabilistic logic learning framework to incorporate both sentence-level and cross-sentence linguistic indicators for indirect supervision. We feed the extracted evidence sentences into existing MRC models and evaluate the end-to-end performance on three challenging multiple-choice MRC datasets: MultiRC, RACE, and DREAM, achieving comparable or better performance than the same models that take as input the full reference document. To the best of our knowledge, this is the first work extracting evidence sentences for multiple-choice MRC.

1 Introduction

Recently, there have been increased interests in machine reading comprehension (MRC). In this work, we mainly focus on multiple-choice MRC (Richardson et al., 2013; Mostafazadeh et al., 2016; Ostermann et al., 2018): given a document and a question, the task aims to select the correct answer option(s) from a small number of answer options associated with this question. Compared to extractive and abstractive MRC tasks (e.g., (Rajpurkar et al., 2016; Kočiskỳ et al., 2018; Reddy et al., 2019)) where most questions can be answered using spans from the reference documents, the majority of answer options cannot be directly extracted from the given texts.

Existing multiple-choice MRC models (Wang et al., 2018b; Radford et al., 2018) take as input the entire reference document and seldom offer any explanation, making interpreting their predictions extremely difficult. It is a natural choice for human readers to use sentences from a given text to explain why they select a certain answer option in reading tests (Bax, 2013). In this paper, as a preliminary attempt, we focus on exacting ***evidence sentences*** that entail or support a question-answer pair from the given reference document.

For extractive MRC tasks, information retrieval techniques can be very strong baselines to extract sentences that contain questions and their answers when questions provide sufficient information, and most questions are factoid and answerable from the content of a single sentence (Lin et al., 2018; Min et al., 2018). However, we face unique challenges to extract evidence sentences for multiple-choice MRC tasks. The correct answer options of a significant number of questions (e.g., 87% questions in RACE (Lai et al., 2017; Sun et al., 2019)) are not extractive, which may require advanced reading skills such as inference over multiple sentences and utilization of prior knowledge (Lai et al., 2017; Khashabi et al., 2018; Ostermann et al., 2018). Besides, the existence of misleading wrong answer options also dramatically increases the difficulty of evidence sentence extraction, especially when a question provides insufficient information. For example, in Figure 1, given the reference document and question *"Which of the following statements is true according to the passage?"*, almost all the tokens in

[*] This work was done when H. W. and K. S. were at Tencent AI Lab, Bellevue, WA.

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 696–707
Hong Kong, China, November 3-4, 2019. ©2019 Association for Computational Linguistics

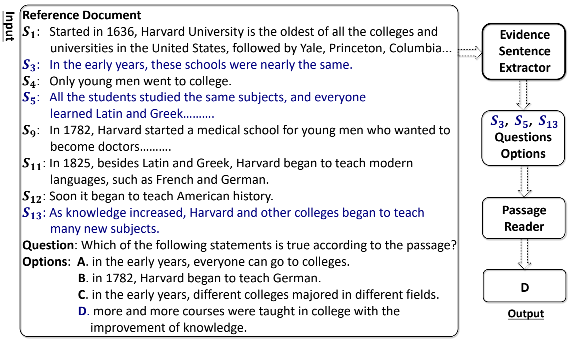
the wrong answer option B *"In 1782, Harvard began to teach German."* appear in the document (i.e., sentence S_9 and S_{11}) while the question gives little useful information for locating answers. Furthermore, we notice that even humans sometimes have difficulty in finding pieces of evidence when the relationship between a question and its correct answer option is implicitly indicated in the document (e.g., *"What is the main idea of this passage?"*). Considering these challenges, we argue that extracting evidence sentences for multiple-choice MRC is at least as difficult as that for extractive MRC or factoid question answering.

Given a question, its associated answer options, and a reference document, we propose a method to extract sentences that can potentially support or explain the (question, correct answer option) pair from the reference document. Due to the lack of ground truth evidence sentences in most multiple-choice MRC tasks, inspired by distant supervision, we first extract *silver standard* evidence sentences based on the lexical features of a question and its correct answer option (Section 2.2), then we use these noisy labels to train an evidence sentence extractor (Section 2.1). To denoise imperfect labels, we also manually design sentence-level and cross-sentence linguistic indicators such as *"adjacent sentences tend to have the same label"* and accommodate all the linguistic indicators with a recently proposed deep probabilistic logic learning framework (Wang and Poon, 2018) for indirect supervision (Section 2.3).

Previous extractive MRC and question answering studies (Min et al., 2018; Lin et al., 2018) indicate that a model should be able to achieve comparable end-to-end performance if it can accurately predict the evidence sentence(s). Inspired by the observation, to indirectly evaluate the quality of the extracted evidence sentences, we only keep the selected sentences as the new reference document for each instance and evaluate the performance of a machine reader (Wang et al., 2018b; Radford et al., 2018) on three challenging multiple-choice MRC datasets: MultiRC (Khashabi et al., 2018), RACE (Lai et al., 2017), and DREAM (Sun et al., 2019). Experimental results show that we can achieve comparable or better performance than the same reader that considers the full context. The comparison between ground truth evidence sentences and automatically selected sentences indicates that there is still room for improvement.

Our primary contributions are as follows: 1) to the best of our knowledge, this is the first work to extract evidence sentences for multiple-choice MRC; 2) we show that it may be a promising direction to leverage various sources of linguistic knowledge for denoising noisy evidence sentence labels. We hope our attempts and observations can encourage the research community to develop more explainable MRC models that simultaneously provide predictions and textual evidence.

2 Method


Input
Evidence Sentence Extractor
S_3, S_5, S_{13} Questions Options
Passage Reader
D
Output

Reference Document
S_1: Started in 1636, Harvard University is the oldest of all the colleges and universities in the United States, followed by Yale, Princeton, Columbia...
S_3: In the early years, these schools were nearly the same.
S_4: Only young men went to college.
S_5: All the students studied the same subjects, and everyone learned Latin and Greek.........
S_9: In 1782, Harvard started a medical school for young men who wanted to become doctors.........
S_{11}: In 1825, besides Latin and Greek, Harvard began to teach modern languages, such as French and German.
S_{12}: Soon it began to teach American history.
S_{13}: As knowledge increased, Harvard and other colleges began to teach many new subjects.
Question: Which of the following statements is true according to the passage?
Options: **A**. in the early years, everyone can go to colleges.
B. in 1782, Harvard began to teach German.
C. in the early years, different colleges majored in different fields.
D. more and more courses were taught in college with the improvement of knowledge.

Figure 1: An overview of our pipeline. The input instance comes from RACE (Lai et al., 2017).

We will present our evidence sentence extractor (Section 2.1) trained on the noisy training data generated by distant supervision (Section 2.2) and denoised by an existing deep probabilistic logic framework that incorporates different kinds of linguistic indicators (Section 2.3). The extractor is followed by an independent neural reader for evaluation. See an overview in Figure 1.

2.1 Evidence Sentence Extractor

We use a multi-layer multi-head transformer (Vaswani et al., 2017) to extract evidence sentences. Let W_w and W_p be the word (subword) and position embeddings, respectively. Let M denote the total number of layers in the transformer. Then, the m-th layer hidden state h^m of a token is given by:

$$h^m = \begin{cases} W_w + W_p & \text{if } m = 0 \\ \mathrm{TB}(h^{m-1}) & \text{if } 1 \leq m \leq M \end{cases} \tag{1}$$

where TB stands for the Transformer Block, which is a standard module that contains MLP, residual connections (He et al., 2016) and LayerNorm (Ba et al., 2016).

Compared to classical transformers, pre-trained transformers such as GPT (Radford et al., 2018) and BERT (Devlin et al., 2018) capture rich world and linguistic knowledge from large-scale external corpora, and significant improvements are obtained by fine-tuning these pre-trained models on a variety of downstream tasks. We follow this promising direction by fine-tuning GPT (Radford et al., 2018) on a target task. Note that the pre-trained transformer in our pipeline can also be easily replaced with other pre-trained models, which however is not the focus of this paper.

We use (X, Y) to denote all training data, (X_i, Y_i) to denote each instance, where X_i is a token sequence, namely, $X_i = \{X_i^1, \ldots, X_i^t\}$ where t equals to the sequence length. For evidence sentence extraction, X_i contains one sentence in a document, a question, and all answer options associated with the question. Y_i indicates the probability that sentence in X_i is selected as an evidence sentence for this question, and $\sum_{i=1}^{N} Y_i = 1$, where N equals to the total number of sentences in a document. GPT takes as input X_i and produces the final hidden state h_i^M of the last token in X_i, which is further fed into a linear layer followed by a softmax layer to generate the probability:

$$P_i = \frac{\exp(W_y h_i^M)}{\sum_{1 \leq i \leq N} \exp(W_y h_i^M)} \qquad (2)$$

where W_y is weight matrix of the output layer. We use Kullback-Leibler divergence loss $\mathrm{KL}(Y||P)$ as the training criterion.

We first apply distant supervision to generate noisy evidence sentence labels (Section 2.2). To denoise the labels, during the training stage, we use deep probabilistic logic learning (DPL) – a general framework for combining indirect supervision strategies by composing probabilistic logic with deep learning (Wang and Poon, 2018). Here we consider both sentence-level and cross-sentence linguistic indicators as indirect supervision strategies (Section 2.3).

As shown in Figure 2, during training, our evidence sentence extractor contains two components: a probabilistic graph containing various sources of indirect supervision used as a supervision module and a fine-tuned GPT used as a prediction module. The two components are connected via a set of latent variables indicating whether each sentence is an evidence sentence or not. We update the model by alternatively optimizing GPT and the probabilistic graph so that they reach an agreement on latent variables. After training, only the fine-tuned GPT is kept to make predictions for a new instance during testing. We provide more details in Appendix A and refer readers to Wang and Poon (2018) for how to apply DPL as a tool in a downstream task such as relation extraction.

2.2 Silver Standard Evidence Generation

Given correct answer options, we use a distant supervision method to generate the *silver standard* evidence sentences.

Inspired by Integer Linear Programming models (ILP) for summarization (Berg-Kirkpatrick et al., 2011; Boudin et al., 2015), we model evidence sentence extraction as a maximum coverage problem and define the value of a selected sentence set as the sum of the weights for the unique words it contains. Formally, let v_i denote the weight of word i, $v_i = 1$ if word i appears in the correct answer option, $v_i = 0.1$ if it appears in the question but not in the correct answer option, and $v_i = 0$ otherwise.[1]

We use binary variables c_i and s_j to indicate the presence of word i and sentence j in the selected sentence set, respectively. $\mathrm{Occ}_{i,j}$ is a binary variable indicating the occurrence of word i in sentence j, l_j denotes the length of sentence j, and L is the predefined maximum number of selected sentences. We formulate the ILP problem as:

$$\max \sum_i v_i c_i \quad s.t. \sum_j s_j \leq L \qquad (3)$$

$$s_j \,\mathrm{Occ}_{ij} \leq c_i, \forall i, j \qquad \sum_j s_j \,\mathrm{Occ}_{ij} \geq c_i, \forall i \qquad (4)$$

$$c_i \in \{0, 1\} \,\forall i \quad s_j \in \{0, 1\} \,\forall j$$

2.3 Linguistic Indicators for Indirect Supervision

To denoise the imperfect labels generated by distant supervision (Section 2.2), as a preliminary attempt, we manually design a small number of sentence-level and cross-sentence linguistic indicators incorporated in DPL for indirect supervision. We briefly introduce them as follows and detail all indicators in Appendix A.3 and implementation details in Section 3.2.

We assume that a sentence is more likely to be an evidence sentence if the sentence and the question have similar meanings, lengths, coherent entity types, same sentiment polarity, or the

[1] We do not observe a significant improvement by tuning parameters v_i on the development set.

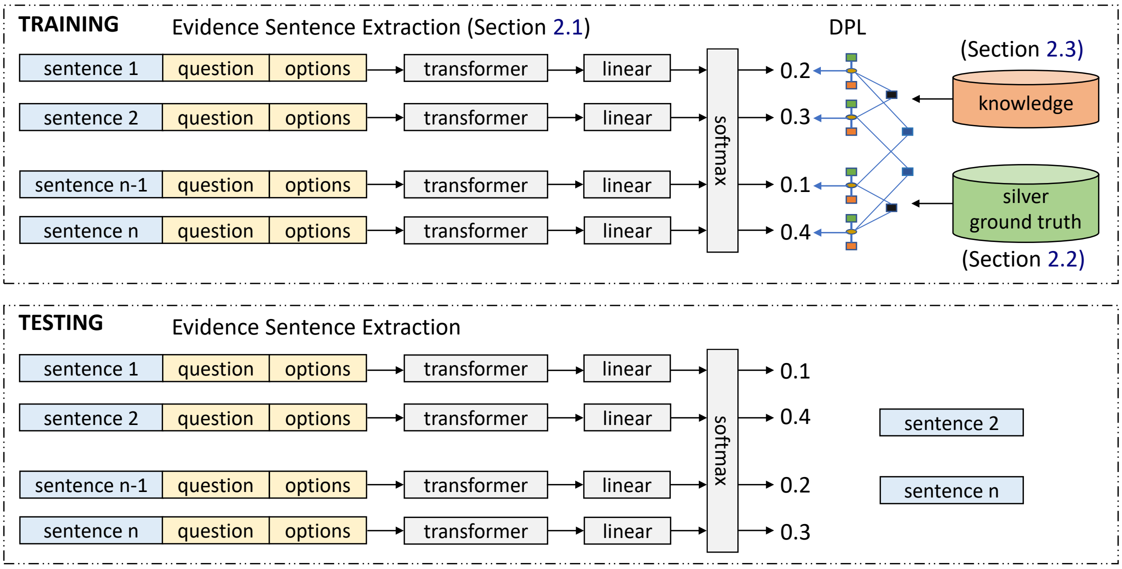


Figure 2: Deep probabilistic logic (DPL) framework for evidence sentence extraction. During testing, we only use trained evidence sentence extractor for prediction.

sentence is true (i.e., entailment) given the question. We assume that a good evidence sentence should be neither too long nor too short (i.e., $5 \leq$ $\#$ of tokens in sentence ≤ 40) considering informativeness and conciseness, and an evidence sentence is more likely to lead to the prediction of the correct answer option (referred as "reward"), which is motivated by our experiments that machine readers take as input the silver (or gold) standard evidence sentences achieve the best performance except for human performance on three multiple-choice machine reading comprehension datasets (Table 2, Table 3, and Table 4 in Section 3). We rely on both lexical features (e.g., lengths and entity types) and semantic features based on pre-trained word/paraphrase embeddings and external knowledge graphs to measure the similarity of meanings. We use existing models or resources for reward calculation, sentiment analysis and natural language inference.

For cross-sentence indicators, we consider that the same set of evidence sentences are less likely to support multiple questions and two evidence sentences that support the same question should be within a certain distance (i.e., evidence sentences for the same question should be within window size 8 (in sentences)). We also assume that adjacent sentences tend to have the same label. We will have more discussions about these assumptions in the data analysis (Section 3.6).

3 Experiments

3.1 Datasets

We use the following three latest multiple-choice machine reading comprehension datasets for evaluation. We show data statistics in Table 1.

MultiRC (Khashabi et al., 2018): MultiRC is a dataset in which questions can only be answered by considering information from multiple sentences. A question may have multiple correct answer options. Reference documents come from seven different domains such as elementary school science and travel guides. For each document, questions and their associated answer options are generated and verified by turkers.

RACE (Lai et al., 2017): RACE is a dataset collected from English language exams designed for middle (RACE-Middle) and high school (RACE-High) students in China, carefully designed by English instructors. The proportion of questions that requires reasoning is 59.2%.

DREAM (Sun et al., 2019): DREAM is a dataset collected from English exams for Chinese language learners. Each instance in DREAM contains a multi-turn multi-party dialogue, and the correct answer option must be inferred from the dialogue context. In particular, a large portion of questions require multi-sentence inference (84%) and/or commonsense knowledge (34%).

Dataset	# of documents			# of questions			Average # of sentences per document
	Train	Dev	Test	Train	Dev	Test	Train + Dev + Test
MultiRC	456	83	332	5,131	953	3,788	14.5 (Train + Dev)
DREAM	3,869	1,288	1,287	6,116	2,040	2,041	8.5
RACE	25,137	1,389	1,407	87,866	4,887	4,934	17.6

Table 1: Statistics of multiple-choice machine reading comprehension and question answering datasets.

3.2 Implementation Details

We use spaCy (Honnibal and Johnson, 2015) for tokenization and named entity tagging. We use the pre-trained transformer (i.e., GPT) released by Radford et al. (2018) with the same preprocessing procedure. When GPT is used as the neural reader, we set training epochs to 4, use eight P40 GPUs for experiments on RACE, and use one GPU for experiments on other datasets. When GPT is used as the evidence sentence extractor, we set batch size 1 per GPU and dropout rate 0.3. We keep other parameters default. Depending on the dataset, training the evidence sentence extractor generally takes several hours.

For DPL, we adopt the toolkit from Wang and Poon (2018). During training, we conduct message passing in DPL iteratively, which usually converges within 5 iterations. We use Vader (Gilbert, 2014) for sentiment analysis and ParaNMT-50M (Wieting and Gimpel, 2018) to calculate the paraphrase similarity between two sentences. We use the knowledge graphs (i.e., triples in ConceptNet v5.6 (Speer and Havasi, 2012; Speer et al., 2017)) to incorporate commonsense knowledge. To calculate the natural language inference probability, we first fine-tune the transformer (Radford et al., 2018) on several tasks, including SNLI (Bowman et al., 2015), SciTail (Khot et al., 2018), MultiNLI (Williams et al., 2018), and QNLI (Wang et al., 2018a).

To calculate the probability that each sentence leads to the correct answer option, we sample a subset of sentences and use them to replace the full context in each instance, and then we feed them into the transformer fine-tuned with instances with full context. If a particular combination of sentences $S = \{s_1, \ldots, s_n\}$ leads to the prediction of the correct answer option, we reward each sentence inside this set with $1/n$. To avoid the combinatorial explosion, we assume evidence sentences lie within window size 3. For another neural reader Co-Matching (Wang et al., 2018b), we use its default parameters. For DREAM and RACE, we set L, the maximum number of silver standard evidence sentences of a question, to 3. For MultiRC, we set L to 5 since many questions have more than 5 ground truth evidence sentences.

3.3 Evaluation on MultiRC

Since its test set is not publicly available, currently we only evaluate our model on the development set (Table 2). The fine-tuned transformer (GPT) baseline, which takes as input the full document, achieves an improvement of 2.2% in macro-average F1 (F1_m) over the previous highest score, 66.5%. If we train our evidence sentence extractor using the ground truth evidence sentences provided by turkers, we can obtain a much higher F1_m 72.3%, even after we remove nearly 66% of sentences in average per document. We can regard this result as the supervised upper bound for our evidence sentence extractor. If we train the evidence sentence extractor with DPL as a supervision module, we get 70.5% in F1_m. The performance gap between 70.5% and 72.3% shows there is still room for improving denoising strategies.

3.4 Evaluation on RACE

As we cannot find any public implementations of recently published independent sentence selectors, we compare our evidence sentence extractor with InferSent released by Conneau et al. (2017) as previous work (Htut et al., 2018) has shown that it outperforms many state-of-the-art sophisticated sentence selectors on a range of tasks. We also investigate the ***portability*** of our evidence sentence extractor by combing it with two neural readers. Besides the fine-tuned GPT baseline, we use Co-Matching (Wang et al., 2018b), another state-of-the-art neural reader on the RACE dataset.

As shown in Table 3, by using the evidence sentences selected by InferSent, we suffer up to a 1.9% drop in accuracy with Co-Matching and up to a 4.2% drop with the fine-tuned GPT. In comparison, by using the sentences extracted by our sentence extractor, which is trained with DPL as a

Approach	F1$_m$	F1$_a$	EM$_0$
All-ones baseline (Khashabi et al., 2018)	61.0	59.9	0.8
Lucene world baseline (Khashabi et al., 2018)	61.8	59.2	1.4
Lucene paragraphs baseline (Khashabi et al., 2018)	64.3	60.0	7.5
Logistic regression (Khashabi et al., 2018)	66.5	63.2	11.8
Full context + Fine-Tuned Transformer (GPT, Radford et al. (2018))	68.7	66.7	11.0
Random 5 sentences + GPT	65.3	63.1	7.2
Top 5 sentences by ESE$_{\text{DS}}$ + GPT	70.2	**68.6**	12.7
Top 5 sentences by ESE$_{\text{DPL}}$ + GPT	**70.5**	67.8	**13.3**
Top 5 sentences by ESE$_{\text{gt}}$ + GPT	**72.3**	**70.1**	**19.2**
Ground truth evidence sentences + GPT	78.1	74.0	28.6
Human Performance (Khashabi et al., 2018)	86.4	83.8	56.6

Table 2: Performance of various settings on the MultiRC development set. We use the fine-tuned GPT as the evidence sentence extractor (ESE) and the neural reader (ESE$_{\text{DS}}$: ESE trained on the silver standard evidence sentences; ESE$_{\text{DPL}}$: ESE trained with DPL as a supervision module; ESE$_{\text{gt}}$: ESE trained using ground truth evidence sentences; F1$_{\text{m}}$ macro-average F1; F1$_{\text{a}}$: micro-average F1; EM$_0$: exact match).

Approach		**Dev**			**Test**	
	Middle	High	All	Middle	High	All
Sliding Window (Richardson et al., 2013; Lai et al., 2017)	-	-	-	37.3	30.4	32.2
Co-Matching (Wang et al., 2018b)	-	-	-	55.8	48.2	50.4
Full context + GPT (Radford et al., 2018)	-	-	-	62.9	57.4	59.0
Full context + GPT	55.6	56.5	56.0	57.5	56.5	56.8
Random 3 sentences + GPT	50.3	51.1	50.9	50.9	49.5	49.9
Top 3 sentences by InferSent (question) + Co-Matching	49.8	48.1	48.5	50.0	45.5	46.8
Top 3 sentences by InferSent (question + all options) + Co-Matching	52.6	49.2	50.1	52.6	46.8	48.5
Top 3 sentences by ESE$_{\text{DS}}$ + Co-Matching	58.1	51.6	53.5	55.6	48.2	50.3
Top 3 sentences by ESE$_{\text{DPL}}$ + Co-Matching	57.5	52.9	54.2	57.5	49.3	51.6
Top 3 sentences by InferSent (question) + GPT	55.0	54.7	54.8	54.6	53.4	53.7
Top 3 sentences by InferSent (question + all options) + GPT	59.2	54.6	55.9	57.2	53.8	54.8
Top 3 sentences by ESE$_{\text{DS}}$ + GPT	62.5	57.7	59.1	64.1	55.4	58.0
Top 3 sentences by ESE$_{\text{DPL}}$ + GPT	63.2	56.9	58.8	**64.3**	56.7	58.9
Top 3 sentences by ESE$_{\text{DS}}$ + full context + GPT	63.4	58.6	60.0	63.7	57.7	59.5
Top 3 sentences by ESE$_{\text{DPL}}$ + full context + GPT	64.2	58.5	60.2	62.4	**58.7**	**59.8**
Silver standard evidence sentences + GPT	73.2	73.9	73.7	74.1	72.3	72.8
Amazon Turker Performance (Lai et al., 2017)	-	-	-	85.1	69.4	73.3
Ceiling Performance (Lai et al., 2017)	-	-	-	95.4	94.2	94.5

Table 3: Accuracy (%) of various settings on the RACE dataset. ESE$_{\text{DS}}$: evidence sentence extractor trained on the silver standard evidence sentences extracted from the rule-based distant supervision method.

supervision module, we observe a much smaller decrease (0.1%) in accuracy with the fine-tuned GPT baseline, and we slightly improve the accuracy with the Co-Matching baseline. For questions in RACE, introducing the content of answer options as additional information for evidence sentence extraction can narrow the accuracy gap, which might be due to the fact that many questions are less informative (Xu et al., 2018). Note that all these results are compared with 59% reported from Radford et al. (2018), if compared with our own replication (56.8%), sentence extractor trained with either DPL or distant supervision leads to gain up to 2.1%.

Since the problems in RACE are designed for human participants that require advanced reading comprehension skills such as the utilization of external world knowledge and in-depth reasoning, even human annotators sometimes have difficulties in locating evidence sentences (Section 3.6). Therefore, ***a limited number of evidence sentences might be insufficient for answering challenging questions***. Instead of removing "non-relevant" sentences, we keep all the sentences in a document while adding a special token before and after the extracted evidence sentences. With DPL as a supervision module, we see an improvement in accuracy of 0.9% (from 58.9% to 59.8%).

For our current supervised upper bound (i.e., assuming we know the correct answer option, we find the silver evidence sentences from rule-based distant supervision and then feed them into the fine-tuned transformer, we get 72.8% in accuracy, which is quite close to the performance of Amazon Turkers. However, it is still much lower than the ceiling performance. To answer questions that require external knowledge, ***it might be a promising direction to retrieve evidence sentences from external resources***, compared to only considering sentences within a reference document for multiple-choice machine reading comprehension tasks.

3.5 Evaluation on DREAM

See Table 4 for results on the DREAM dataset. The fine-tuned GPT baseline, which taks as input the full document, achieves 55.1% in accuracy on the test set. If we train our evidence sentence extractor with DPL as a supervision module and feed the extracted evidence sentences to the fine-tuned GPT, we get test accuracy 57.7%. Similarly, if we train the evidence sentence extractor only with silver standard evidence sentences extracted from the rule-based distant supervision method, we obtain test accuracy 56.3%, i.e., 1.4% lower than that with full supervision. Experiments demonstrate the effectiveness of our evidence sentence extractor with denoising strategy, and the usefulness of evidence sentences for dialogue-based machine reading comprehension tasks in which reference documents are less formal compared to those in RACE and MultiRC.

Approach	Dev	Test
Full context + GPT† (Sun et al., 2019)	55.9	55.5
Full context + GPT	55.1	55.1
Top 3 sentences by ESE$_{\text{silver-gt}}$ + GPT	50.1	50.4
Top 3 sentences by ESE$_{\text{DS}}$ + GPT	55.1	56.3
Top 3 sentences by ESE$_{\text{DPL}}$ + GPT	57.3	**57.7**
Silver standard evidence sentences + GPT	60.5	59.8
Human Performance†	93.9	95.5

Table 4: Performance in accuracy (%) on the DREAM dataset (Results marked with † are taken from Sun et al. (2019); ESE$_{\text{silver-gt}}$: ESE trained using silver standard evidence sentences).

3.6 Human Evaluation

Extracted evidence sentences, which help neural readers to find correct answers, may still fail to convince human readers. Thus we evaluate the quality of extracted evidence sentences based on human annotations (Table 5).

Dataset	Silver Sentences	Sentences by ESE$_{\text{DPL}}$
RACE-M	59.9	57.5
MultiRC	53.0	60.8

Table 5: Macro-average F1 compared with human annotated evidence sentences on the dev set (silver sentences: evidence sentences extracted by ILP (Section 2.2); sentences by ESE$_{\text{DPL}}$: evidence sentences extracted by ESE trained on silver stand ground truth, GT: ground truth evidence sentences).

MultiRC: Even trained using the noisy labels, we achieve a macro-average F1 score 60.8% on MultiRC, indicating the learning and generalization capabilities of our evidence sentence extractor, compared to 53.0%, achieved by using the noisy silver standard evidence sentences guided by correct answer options.

RACE: Since RACE does not provide the ground truth evidence sentences, to get the ground truth evidence sentences, two human annotators annotate 500 questions from the RACE-Middle development set.[2] The Cohen's kappa coefficient between two annotations is 0.87. For negation questions which include negation words (e.g., *"Which statement is not true according to the passage?"*), we have two annotation strategies: we can either find sentences that can directly imply the correct answer option; or the sentences that support the wrong answer options. During annotation, for each question, we use the strategy that leads to fewer evidence sentences.

We find that even humans have troubles in locating evidence sentences when the relationship between a question and its correct answer option is implicitly implied. For example, a significant number of questions require the understanding of the entire document (e.g., *"what's the best title of this passage"* and *"this passage mainly tells us that _"*) and/or external knowledge (e.g., *"the writer begins with the four questions in order to _"*, *"The passage is probably from _"* , and *"If the writer continues the article, he would most likely write about_"*). For 10.8% of total questions, at least one annotator leave the slot blank due to the challenges mentioned above. 65.2% of total questions contain at least two evidence sentences, and

[2] Annotations are available at https://github.com/nlpdata/evidence.

70.9% of these questions contain at least one adjacent sentence pair in their evidence sentences, which may provide evidence to support our assumption *adjacent sentences tend to have the same label* in Section 2.3.

The average and the maximum number of evidence sentences for the remaining questions is 2.1 and 8, respectively. The average number of evidence sentences in the full RACE dataset should be higher since questions in RACE-High are more difficult (Lai et al., 2017), and we ignore 10.8% of the total questions that require the understanding of the whole context.

3.7 Error Analysis

We analyze the predicted evidence sentences for instances in RACE for error analysis. Tough with a high macro-average recall (67.9%), it is likely that our method extracts sentences that support distractors. For example, to answer the question *"You lost your keys. You may call _"*, our system mistakenly extracts sentences *"Please call 5016666"* that support one of the distractors and adjacent to the correct evidence sentences *"Found a set of keys. Please call Jane at 5019999."* in the given document. We may need linguistic constraints or indicators to filter out irrelevant selected sentences instead of simply setting a hard length constraint such as 5 for all instances in a dataset.

Besides, it is possible that there is no clear sentence in the document for justifying the correctness of the correct answer. For example, to answer the question *"What does "figure out" mean ?"*, neither *"find out"* nor the correct answer option appears in the given document as this question mainly assesses the vocabulary acquisition of human readers. Therefore, all the extracted sentences (e.g., *"sometimes... sometimes I feel lonely, like I'm by myself with no one here."*, *"sometimes I feel excited, like I have some news I have to share!"*) by our methods are inappropriate. A possible solution is to predict whether a question is answerable following previous work (e.g., (Hu et al., 2019)) on addressing unanswerable questions in extractive machine reading comprehension tasks such as SQuAD (Rajpurkar et al., 2018) before to extract the evidence sentences for this question.

4 Related Work

4.1 Sentence Selection for Machine Reading Comprehension and Fact Verification

Previous studies investigate paragraph retrieval for factoid question answering (Chen et al., 2017; Wang et al., 2018c; Choi et al., 2017; Lin et al., 2018), sentence selection for machine reading comprehension (Hewlett et al., 2017; Min et al., 2018), and fact verification (Yin and Roth, 2018; Hanselowski et al., 2018). In these tasks, most of the factual questions/claims provide sufficient clues for identifying relevant sentences, thus often information retrieval combined with filters can serve as a very strong baseline. For example, in the FEVER dataset (Thorne et al., 2018), only 16.8% of claims require composition of multiple evidence sentences. For some of the cloze-style machine reading comprehension tasks such as CBT (Hill et al., 2016), Kaushik and Lipton (2018) demonstrate that for some models, comparable performance can be achieved by considering only the last sentence that usually contains the answer. Different from above work, we exploit information in answer options and use various indirect supervision to train our evidence sentence extractor, and previous work can actually be a regarded as a special case for our pipeline. Compared to Lin et al. (2018), we leverage rich linguistic knowledge for denoising imperfect labels.

Several work also investigate content selection at the token level (Yu et al., 2017; Seo et al., 2018), in which some tokens are automatically skipped by neural models. However, they do not utilize any linguistic knowledge, and a set of discontinuous tokens has limited explanation capability.

4.2 Machine Reading Comprehension with External Linguistic Knowledge

Linguistic knowledge such as coreference resolution, frame semantics, and discourse relations is widely used to improve machine comprehension (Wang et al., 2015; Sachan et al., 2015; Narasimhan and Barzilay, 2015; Sun et al., 2018) especially when there are only hundreds of documents available in a dataset such as MCTest (Richardson et al., 2013). Along with the creation of large-scale reading comprehension datasets, recent machine reading comprehension models rely on end-to-end neural models, and it primarily uses word embeddings as input. However, Wang et al. (2016); Dhingra et al. (2017,

2018) show that existing neural models do not fully take advantage of the linguistic knowledge, which is still valuable for machine reading comprehension. Besides widely used lexical features such as part-of-speech tags and named entity types (Wang et al., 2016; Liu et al., 2017; Dhingra et al., 2017, 2018), we consider more diverse types of external knowledge for performance improvements. Moreover, we accommodate external knowledge with probabilistic logic to potentially improve the interpretability of MRC models instead of using external knowledge as additional features.

4.3 Explainable Machine Reading Comprehension and Question Answering

To improve the interpretability of question answering, previous work utilize interpretable internal representations (Palangi et al., 2017) or reasoning networks that employ a hop-by-hop reasoning process dynamically (Zhou et al., 2018). A research line focuses on visualizing the whole derivation process from the natural language utterance to the final answer for question answering over knowledge bases (Abujabal et al., 2017) or scientific word algebra problems (Ling et al., 2017). Jansen et al. (2016) extract explanations that describe the inference needed for elementary science questions (e.g., *"What form of energy causes an ice cube to melt"*). In comparison, the derivation sequence is less apparent for open-domain questions, especially when they require external domain knowledge or multiple-sentence reasoning. To improve explainability, we can also check the attention map learned by neural readers (Wang et al., 2016), however, attention map is learned in end-to-end fashion, which is different from our work.

A similar work proposed by Sharp et al. (2017) also uses distant supervision to learn how to extract informative justifications. However, their experiments are primarily designed for factoid question answering, in which it is relatively easy to extract justifications since most questions are informative. In comparison, we focus on multi-choice MRC that requires deep understanding, and we pay particular attention to denoising strategies.

5 Conclusions

We focus on extracting evidence sentences for multiple-choice MRC tasks, which has not been studied before. We propose to apply distant supervision to noisy labels and apply a deep probabilistic logic framework that incorporates linguistic indicators for denoising noisy labels during training. To indirectly evaluate the quality of the extracted evidence sentences, we feed extracted evidence sentences as input to two existing neural readers. Experimental results show that we can achieve comparable or better performance on three multiple-choice MRC datasets, in comparison with the same readers taking as input the entire document. However, there still exist significant differences between the predicted sentences and ground truth sentences selected by humans, indicating the room for further improvements.

Acknowledgments

We thank the anonymous reviewers for their encouraging and helpful feedback.

References

Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2017. QUINT: Interpretable question answering over knowledge bases. In *Proceedings of the EMNLP (System Demonstrations)*, pages 61–66, Copenhagen, Denmark.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint*, stat.ML/1607.06450v1.

Stephen Bax. 2013. The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, 30(4):441–465.

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of the ACL*, pages 481–490, Portland, OR.

Lidong Bing, Sneha Chaudhari, Richard Wang, and William Cohen. 2015. Improving distant supervision for information extraction using label propagation through lists. In *Proceedings of the EMNLP*, pages 524–529, Lisbon, Portugal.

Florian Boudin, Hugo Mougard, and Benoit Favre. 2015. Concept-based summarization using integer linear programming: From concept pruning to multiple optimal solutions. In *Proceedings of the EMNLP*, pages 17–21, Lisbon, Portugal.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the EMNLP*, pages 632–642, Lisbon, Portuga.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the ACL*, pages 1870–1879, Vancouver, Canada.

Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit, Illia Polosukhin, Alexandre Lacoste, and Jonathan Berant. 2017. Coarse-to-fine question answering for long documents. In *Proceedings of the ACL*, pages 209–220, Vancouver, Canada.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the EMNLP*, pages 670–680, Copenhagen, Denmark.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL*, pages 4171–4186, Minneapolis, MN.

Bhuwan Dhingra, Qiao Jin, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2018. Neural models for reasoning over multiple mentions using coreference. In *Proceedings of the NAACL-HLT*, pages 42–48, New Orleans, LA.

Bhuwan Dhingra, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2017. Linguistic knowledge as memory for recurrent neural networks. *arXiv preprint*, cs.CL/arXiv:1703.02620v1.

CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the ICWSM*, pages 216–225, Qubec, Canada.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. Ukp-athene: Multi-sentence textual entailment for claim verification. *arXiv preprint*, cs.IR/1809.01479v2.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the CVPR*, pages 770–778, Las Vegas, NV.

Daniel Hewlett, Llion Jones, Alexandre Lacoste, et al. 2017. Accurate supervised and semi-supervised machine reading for long documents. In *Proceedings of the EMNLP*, pages 2011–2020, Copenhagen, Denmark.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *Proceedings of the ICLR*, Caribe Hilton, Puerto Rico.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the EMNLP*, pages 1373–1378, Lisbon, Portugal.

Phu Mon Htut, Samuel Bowman, and Kyunghyun Cho. 2018. Training a ranking function for open-domain question answering. In *Proceedings of the NAACL-HLT (Student Research Workshop)*, pages 120–127, New Orleans, LA.

Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huang, Nan Yang, and Dongsheng Li. 2019. Read+ verify: Machine reading comprehension with unanswerable questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6529–6537, Honolulu, HI.

Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 2016. What's in an explanation? Characterizing knowledge and inference requirements for elementary science exams. In *Proceedings of the COLING*, pages 2956–2965, Osaka, Japan.

Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the EMNLP*, pages 5010–5015, Brussels, Belgium.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the NAACL-HLT*, pages 252–262, New Orleans, LA.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI*, pages 5189–5197, New Orleans, LA.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gáabor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association of Computational Linguistics*, 6:317–328.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations. In *Proceedings of the EMNLP*, pages 785–794, Copenhagen, Denmark.

Yankai Lin, Haozhe Ji, Zhiyuan Liu, and Maosong Sun. 2018. Denoising distantly supervised open-domain question answering. In *Proceedings of the ACL*, pages 1736–1745, Melbourne, Australia.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the ACL*, pages 158–167, Vancouver, Canada.

Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2017. Stochastic answer networks for machine reading comprehension. In *Proceedings of the ACL*, pages 1694–1704, Melbourne, Australia.

Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. 2018. Efficient and robust question answering from minimal context over documents. In *Proceedings of the ACL*, pages 1725–1735, Melbourne, Australia.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. In *Proceedings of the NAACL-HLT*, pages 839–849, San Diego, CA.

Karthik Narasimhan and Regina Barzilay. 2015. Machine comprehension with discourse relations. In *Proceedings of the ACL*, pages 1253–1262, Beijing, China.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 Task 11: Machine comprehension using commonsense knowledge. In *Proceedings of the SemEval*, pages 747–757, New Orleans, LA.

Hamid Palangi, Paul Smolensky, Xiaodong He, and Li Deng. 2017. Question-answering with grammatically-interpretable representations. *arXiv preprint*, cs.CL/1705.08432v2.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. In *Preprint*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of th ACL*, pages 784–789, Melbourne, Australia.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the EMNLP*, pages 2383–2392, Austin, TX.

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

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the EMNLP*, pages 193–203, Seattle, WA.

Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. *Machine learning*, 62(1-2):107–136.

Mrinmaya Sachan, Kumar Dubey, Eric Xing, and Matthew Richardson. 2015. Learning answer-entailing structures for machine comprehension. In *Proceedings of the ACL*, pages 239–249, Beijing, China.

Minjoon Seo, Sewon Min, Ali Farhadi, and Hannaneh Hajishirzi. 2018. Neural speed reading via Skim-RNN. In *Proceedings of the ICLR*, New Orleans, LA.

Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Marco A Valenzuela-Escárcega, Peter Clark, and Michael Hammond. 2017. Tell me why: Using question answering as distant supervision for answer justification. In *Proceedings of the CoNLL*, pages 69–79, Vancouver, Canada.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI*, pages 4444–4451, San Francisco, CA.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in ConceptNet 5. In *Proceedings of the LREC*, pages 3679–3686, Istanbul, Turkey.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *Transactions of the Association of Computational Linguistics*, 7:217–231.

Yawei Sun, Gong Cheng, and Yuzhong Qu. 2018. Reading comprehension with graph-based temporal-casual reasoning. In *Proceedings of the COLING*, pages 806–817, Santa Fe, NM.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the NAACL-HLT*, pages 809–819, New Orleans, LA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the NIPS*, pages 5998–6008, Long Beach, CA.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint*, cs.CL/1804.07461v1.

Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2015. Machine comprehension with syntax, frames, and semantics. In *Proceedings of the ACL*, pages 700–706, Beijing, China.

Hai Wang, Takeshi Onishi, Kevin Gimpel, and David McAllester. 2016. Emergent predication structure in hidden state vectors of neural readers. In *Proceedings of the Repl4NLP*, pages 26–36, Vancouver, Canada.

Hai Wang and Hoifung Poon. 2018. Deep probabilistic logic: A unifying framework for indirect supervision. In *Proceedings of the EMNLP*, pages 1891–1902, Brussels, Belgium.

Shuohang Wang, Mo Yu, Shiyu Chang, and Jing Jiang. 2018b. A co-matching model for multi-choice reading comprehension. In *Proceedings of the ACL*, pages 1–6, Melbourne, Australia.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerald Tesauro, Bowen Zhou, and Jing Jiang. 2018c. R^3: Reinforced reader-ranker for open-domain question answering. In *Proceedings of the AAAI*, pages 5981–5988, New Orleans, LA.

John Wieting and Kevin Gimpel. 2018. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the ACL*, pages 451–462, Melbourne, Australia.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the NAACL*, pages 1112–1122, New Orleans, LA.

Yichong Xu, Jingjing Liu, Jianfeng Gao, Yelong Shen, and Xiaodong Liu. 2018. Dynamic fusion networks for machine reading comprehension. *arXiv preprint*, cs.CL/1711.04964v2.

Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A two-wing optimization strategy for evidential claim verification. In *Proceedings of the EMNLP*, pages 105–114, Brussels, Belgium.

Adams Wei Yu, Hongrae Lee, and Quoc Le. 2017. Learning to skim text. In *Proceedings of the ACL*, pages 1880–1890, Vancouver, Canada.

Mantong Zhou, Minlie Huang, and Xiaoyan Zhu. 2018. An interpretable reasoning network for multi-relation question answering. In *Proceedings of the COLING*, pages 2010–2022, Santa Fe, NM.