DREAM: A Challenge Data Set and Models for Dialogue-Based Reading Comprehension

Kai Sun[♠*] **Dian Yu**[♡] **Jianshu Chen**[♡] **Dong Yu**[♡] **Yejin Choi**[♢,♣] **Claire Cardie**[♠]

[♠]Cornell University, Ithaca, NY, USA
[♡]Tencent AI Lab, Bellevue, WA, USA
[♢]University of Washington, Seattle, WA, USA
[♣]Allen Institute for Artificial Intelligence, Seattle, WA, USA
ks985@cornell.edu {yudian,jianshuchen,dyu}@tencent.com
yejin@cs.washington.edu cardie@cs.cornell.edu

Abstract

We present DREAM, the first dialogue-based multiple-choice reading comprehension data set. Collected from English as a Foreign Language examinations designed by human experts to evaluate the comprehension level of Chinese learners of English, our data set contains 10,197 multiple-choice questions for 6,444 dialogues. In contrast to existing reading comprehension data sets, DREAM is the first to focus on in-depth multi-turn multi-party dialogue understanding. DREAM is likely to present significant challenges for existing reading comprehension systems: 84% of answers are non-extractive, 85% of questions require reasoning beyond a single sentence, and 34% of questions also involve commonsense knowledge.

We apply several popular neural reading comprehension models that primarily exploit surface information within the text and find them to, at best, just barely outperform a rule-based approach. We next investigate the effects of incorporating dialogue structure and different kinds of general world knowledge into both rule-based and (neural and non-neural) machine learning-based reading comprehension models. Experimental results on the DREAM data set show the effectiveness of dialogue structure and general world knowledge. DREAM is available at https://dataset.org/dream/.

1 Introduction

Recently a significant amount of research has focused on the construction of large-scale multiple-choice (Lai et al., 2017; Khashabi et al., 2018; Ostermann et al., 2018) and extractive (Hermann et al., 2015; Hill et al., 2016; Rajpurkar et al., 2016; Trischler et al., 2017) reading comprehension data sets (Section 2). Source documents in these data sets have generally been drawn from formal written texts such as news, fiction, and Wikipedia articles, which are commonly considered well-written, accurate, and neutral.

With the goal of advancing research in machine reading comprehension and facilitating dialogue understanding, we construct and present DREAM — the first multiple-choice **D**ialogue-based **REA**ding comprehension exa**M**ination data set. We collect 10,197 questions for 6,444 multi-turn multi-party dialogues from English language exams, which are carefully designed by educational experts (e.g., English teachers) to assess the comprehension level of Chinese learners of English. Each question is associated with three answer options, exactly one of which is correct. (See Table 1 for an example.) DREAM covers a variety of topics and scenarios in daily life such as conversations on the street, on the phone, in a classroom or library, at the airport or the office or a shop (Section 3).

Based on our analysis of DREAM, we argue that dialogue-based reading comprehension is at least as difficult as existing non-conversational counterparts. In particular, answering 34% of DREAM questions requires unspoken commonsense knowledge, for example, unspoken scene information. This might be due to the nature of dialogues: For efficient oral communication, people rarely state obvious explicit world knowledge (Forbes and Choi, 2017) such as *"Christmas Day is celebrated on December 25th."* Understanding

[*]This work was done when K. S. was an intern at the Tencent AI Lab, Bellevue, WA.

Transactions of the Association for Computational Linguistics, vol. 7, pp. 217–231, 2019. Action Editor: Jing Jiang.
Submission batch: 9/2018; Revision batch: 12/2018; Final submission: 2/2019; Published 4/2019.

Dialogue 1 (D1)	
W:	Tom, look at your shoes. How dirty they are! You must clean them.
M:	Oh, mum, I just cleaned them yesterday.
W:	They are dirty now. You must clean them again.
M:	I do not want to clean them today. Even if I clean them today, they will get dirty again tomorrow.
W:	All right, then.
M:	Mum, give me something to eat, please.
W:	You had your breakfast in the morning, Tom, and you had lunch at school.
M:	I am hungry again.
W:	Oh, hungry? But if I give you something to eat today, you will be hungry again tomorrow.
Q1	Why did the woman say that she wouldn't give him anything to eat?
A.	Because his mother wants to correct his bad habit.⋆
B.	Because he had lunch at school.
C.	Because his mother wants to leave him hungry.

Table 1: A sample DREAM problem that requires general world knowledge (⋆: the correct answer option).

the social implications of an utterance as well as inferring a speaker's intentions is also regularly required for answering dialogue-based questions. The dialogue content in Table 1, for example, is itself insufficient for readers to recognize the intention of the female speaker (W) in the first question (Q1). However, world knowledge is rarely considered in state-of-the-art reading comprehension models (Tay et al., 2018; Wang et al., 2018b).

Moreover, dialogue-based questions can cover information imparted across multiple turns involving multiple speakers. In DREAM, approximately 85% of questions can only be answered by considering the information from multiple sentences. For example, to answer Q1 in Table 3 later in the paper regarding the date of birth of the male speaker (M), the supporting sentences (in bold) include *"You know, tomorrow is Christmas Day"* from the female speaker and *". . . I am more than excited about my birthday, which will come in two days"* from the male speaker. Compared with "multiple-sentence questions" in traditional reading comprehension data sets, DREAM further requires an understanding of the turn-based structure of dialogue—for example, for aligning utterances with their corresponding speakers.

As only 16% of correct answer options are text spans from the source documents, we primarily explore rule-based methods and state-of-the-art neural models designed for multiple-choice reading comprehension (Section 4). We find first that neural models designed for non–dialogue-based reading comprehension (Chen et al., 2016; Dhingra et al., 2017; Wang et al., 2018b) do not fare well: The highest achieved accuracy is 45.5%, only slightly better than the accuracy (44.6%) of a simple lexical baseline (Richardson et al., 2013). For the most part, these models fundamentally exploit only surface-level information from the source documents. Considering the above-mentioned challenges, however, we hypothesize that incorporating general world knowledge and aspects of the dialogue structure would allow a better understanding of the dialogues. As a result, we modify our baseline systems to include (1) general world knowledge in the form of such as ConceptNet relations (Speer et al., 2017) and a pre-trained language model (Radford et al., 2018), and (2) speaker information for each utterance. Experiments show the effectiveness of these factors on the lexical baselines as well as neural and non-neural machine learning approaches: We acquire up to 11.9% absolute gain in accuracy compared with the highest performance achieved by the state-of-the-art reading comprehension model (Wang et al., 2018b), which mainly relies on explicit surface-level information in the text (Section 5).

Finally, we see a significant gap between the best automated approach (59.5%) and human ceiling performance (98.6%) on the DREAM data set. This provides yet additional evidence that dialogue-based reading comprehension is a very challenging task. We hope that it also inspires the research community to develop methods for the dialogue-based reading comprehension task.

2 Related Work

We divide reading comprehension data sets into three categories based on the types of answers: extractive, abstractive, and multiple choice.

2.1 Extractive and Abstractive Data Sets

In recent years, we have seen increased interest in large-scale cloze/span-based reading comprehension

	SQuAD	NarrativeQA	CoQA	RACE	DREAM (this work)
Answer type	extractive	abstractive	abstractive	multiple-choice	multiple-choice
Source document type	written text	written text	written text	written text	dialogue
# of source documents	536	1,572	8,399	27,933	6,444
Average answer length	3.2	4.7	2.7	5.3	5.3
Extractive (%)	100.0	73.6	66.8	13.0	16.3
Abstractive (%)	0.0	26.4	33.2	87.0	83.7

Table 2: Distribution of answer (or correct answer option) types in three kinds of reading comprehension data sets. Statistics of other data sets come from Reddy et al. (2018), Kočiskỳ et al. (2018), and Lai et al. (2017).

data set construction (Hermann et al., 2015; Hill et al., 2016; Onishi et al., 2016; Rajpurkar et al., 2016; Bajgar et al., 2016; Nguyen et al., 2016; Trischler et al., 2017; Joshi et al., 2017; Choi et al., 2018). We regard them as extractive since candidate answers are usually short spans from source documents. State-of-the-art neural models with attention mechanisms already achieve very high performance based on local lexical information. Recently researchers work on the construction of spoken span-based data sets (Lee et al., 2018; Li et al., 2018) by applying text-to-speech technologies or recruiting human speakers based on formal written document-based data sets such as **SQuAD** (Rajpurkar et al., 2016). Some span-based conversation data sets are constructed from a relatively small size of dialogues from television shows (Chen and Choi, 2016; Ma et al., 2018).

Considering the limitations in extractive data sets, answers in abstractive data sets such as MS MARCO (Nguyen et al., 2016), SearchQA (Dunn et al., 2017), and **NarrativeQA** (Kočiskỳ et al., 2018) are human-crowdsourced based on source documents or summaries. Concurrently, there is a growing interest in conversational reading comprehension such as **CoQA** (Reddy et al., 2018). Because annotators tend to copy spans as answers (Reddy et al., 2018), the majority of answers are still extractive in these data sets (Table 2). Compared to the data sets mentioned above, most of the correct answer options (**83.7%**) in DREAM are free-form text.

2.2 Multiple-Choice Data Sets

We primarily discuss the multiple-choice data sets, in which answer options are not restricted to extractive text spans in the given document. Instead, most of the correct answer options are abstractive (Table 2). Multiple-choice data sets involve extensive human involvement for problem generation during crowdsourcing (i.e., questions, correct answer option, and distractors). Besides surface matching, a significant portion of questions require multiple-sentence reasoning and external knowledge (Richardson et al., 2013; Mostafazadeh et al., 2016; Khashabi et al., 2018; Ostermann et al., 2018).

Besides crowdsourcing, some data sets are collected from examinations designed by educational experts (Penas et al., 2014; Shibuki et al., 2014; Tseng et al., 2016; Clark et al., 2016; Lai et al., 2017; Mihaylov et al., 2018), which aim to test human examinees. There are various types of complicated questions such as math word problems, summarization, logical reasoning, and sentiment analysis. Because we can adopt more objective evaluation criteria such as accuracy, these questions are usually easy to grade. Besides, questions from examinations are generally clean and high-quality. Therefore, human performance ceiling on this kind of data set is much higher (e.g., 94.5% on **RACE** [Lai et al., 2017] and 98.6% on DREAM in accuracy) than that of data sets built by crowdsourcing.

In comparison, we present the first multiple-choice **dialogue-based** data set from examinations that contains a large percentage of questions that require multiple sentence inference. To the best of our knowledge, DREAM also contains the largest number of questions involving **commonsense reasoning** compared with other examination data sets.

3 Data

In this section, we describe how we construct DREAM (Section 3.1) and provide a detailed analysis of this data set (Section 3.2).

3.1 Collection Methodology

We collect dialogue-based comprehension problems from a variety of English language exams

Dialogue 2 (D2)
W: Hey, Mike. Where have you been? I didn't see you around these days?
M: I was hiding in my office. My boss gave me loads of work to do, and I tried to finish it before my birthday. Anyway, I am done now. Thank goodness! How is everything going with you?
W: I'm quite well. **You know, tomorrow is Christmas Day.** Do you have any plans?
M: **Well, to tell you the truth, I am more than excited about my birthday, which will come in two days.** I am going to visit my parents-in-law with my wife.
W: Wow, sounds great.
M: Definitely! This is my first time to spend my birthday with them.
W: Do they live far away from here?
M: A little bit. We planned to take the train, but considering the travel peak, my wife strongly suggested that we go to the airport right after we finish our work this afternoon. How about you? What's your holiday plan?
W: Well, our situations are just the opposite. My parents-in-law will come to my house, and they wish to stay at home and have a quiet Christmas Day. So I have to call my friends to cancel our party that will be held at my house.
M: You'll experience a quite different and lovely holiday. Enjoy your Christmas!
W: Thanks, the same to you!
Q1 What is the date of the man's birthday?
A. 25th, December.
B. 26th, December.⋆
C. 27th, December.
Q2 How will the man go to his wife's parents' home?
A. By train.
B. By bus.
C. By plane.⋆
Q3 What is the probable relationship between the two speakers?
A. Husband and wife.
B. Friends.⋆
C. Parent-in-law and son-in-law.

Table 3: A complete sample DREAM problem (⋆: the correct answer option).

(including practice exams) such as National College Entrance Examination, College English Test, and Public English Test,[1] which are designed by human experts to assess either the listening or reading comprehension level of Chinese English learners in high schools and colleges (for individuals aged 12–22 years). All the problems in DREAM are freely accessible online for public usage. Each problem consists of a dialogue and a series of multiple-choice questions. To ensure every question is associated with exactly three answer options, we drop wrong answer option(s) randomly for questions with more than three options. We remove duplicate problems and randomly split the data at the problem level, with 60% train, 20% development, and 20% test.

[1]We list all the Web sites used for data collection in the released data set.

Metric	Value
# of answer options per question	3
# of turns	30,183
Avg./Max. # of questions per dialogue	1.6 / 10
Avg./Max. # of speakers per dialogue	2.0 / 7
Avg./Max. # of turns per dialogue	4.7 / 48
Avg./Max. option length (in tokens)	5.3 / 21
Avg./Max. question length (in tokens)	8.6 / 24
Avg./Max. dialogue length (in tokens)	85.9 / 1,290
vocabulary size	13,037

Table 4: The overall statistics of DREAM. A turn is defined as an uninterrupted stream of speech from one speaker in a dialogue.

	Train	Dev	Test	All
# of dialogues	3,869	1,288	1,287	6,444
# of questions	6,116	2,040	2,041	10,197

Table 5: The separation of the training, development, and test sets in DREAM.

3.2 Data Analysis

We summarize the statistics of DREAM in Table 4 and data split in Table 5. Compared with existing data sets built from formal written texts, the vocabulary size is relatively small since spoken English by its nature makes greater use of high-frequency words and needs a smaller vocabulary for efficient real-time communication (Nation, 2006).

We categorize questions into two main categories according to the types of knowledge required to answer them: *matching* and *reasoning*.

- **Matching** A question is entailed or paraphrased by exactly one sentence in a dialogue. The answer can be extracted from the same sentence. For example, we can easily verify the correctness of the question-answer pair (*"What kind of room does the man want to rent?"*, *"A two-bedroom apartment."*)

based on the sentence *"**M**: I'm interested in renting a two-bedroom apartment."* This category is further divided into two categories, *word matching* and *paraphrasing,* in previous work (Chen et al., 2016; Trischler et al., 2017).

- **Reasoning** Questions that cannot be answered by the surface meaning of a single sentence belong to this category. We further define four subcategories as follows.
  - **Summary** Answering this kind of questions requires the whole picture of a dialogue, such as the topic of a dialogue and the relation between speakers (e.g., D2-Q3 in Table 3). Under this category, questions such as *"What are the two speakers talking about?"* and *"What are the speakers probably doing?"* are frequently asked.
  - **Logic** We require logical reasoning to answer questions in this category. We usually need to identify logically implied relations among multiple sentences in a dialogue. To reduce the ambiguity during the annotation, we regard a question that can only be solved by considering the content of multiple sentences and does not belong to the *summary* subcategory that involves all the sentences in a dialogue as a *logic* question. Following this definition, both D2-Q1 and D2-Q2 in Table 3 belong to this category.
  - **Arithmetic** Inferring the answer requires arithmetic knowledge (e.g., D2-Q1 in Table 3 requires $25 - 1 + 2 = 26$).
  - **Commonsense** To answer questions under this subcategory, besides the textual information in the dialogue, we also require external commonsense knowledge that cannot be obtained from the dialogue. For instance, all questions in Table 3 fall under this category. D2-Q1 and D2-Q2 in Table 3 belong to both *logic* and *commonsense* since they require multiple sentences as well as commonsense knowledge for question answering. There exist multiple types of commonsense knowledge in DREAM such as the well-known properties of a highly recognizable entity (e.g., D2-Q1 in Table 3), the prominent relationship-between two speakers (e.g., D2-Q3 in Table 3), the knowledge of or shared by a particular culture (e.g., when a speaker says *"Cola? I think it tastes like medicine."*, she/he probably means *"I don't like cola."*), and the cause-effect relation between events (e.g., D1-Q1 in Table 1). We refer readers to LoBue and Yates (2011) for detailed definitions.

Question Type	Dev	Test	Dev + Test
Matching	13.0	10.3	11.7
Reasoning	87.0	89.7	88.3
Summary	8.4	15.9	12.1
Logic	74.5	70.4	72.5
Arithmetic	5.1	3.6	4.4
Commonsense	31.5	35.9	33.7
Single sentence	17.1	13.7	15.4
Multiple sentences	82.9	86.3	84.6

Table 6: Distribution (%) of question types.

Table 6 shows the question type distribution labeled by two human annotators on 25% questions randomly sampled from the development and test sets. Besides the previously defined question categories, we also report the percentage of questions that require reasoning over multiple sentences (i.e., *summary* or *logic* questions) and the percentage of questions that require the surface-level understanding or commonsense/math knowledge based on the content of a single sentence. As a question can belong to multiple reasoning subcategories, the summation of the percentage of reasoning subcategories is not equal to the percentage of reasoning. The Cohen's kappa coefficient is 0.67 on the development set and 0.68 on the test set.

Dialogues in DREAM are generally clean and mostly error-free because they are carefully designed by educational experts. However, it is not guaranteed that each dialogue is written or proofread by a native speaker. Besides, dialogues tend to be more proper and less informal for exam purposes. To have a rough estimation of the quality of dialogues in DREAM and the differences between these dialogues and more casual ones in movies or television shows, we run a proofreading tool—Grammarly[2]—on all the dialogues from the annotated 25% instances of the development set and the same size (20.7k tokens) of dialogues from *Friends*, a famous American television show

[2]https://app.grammarly.com.

Metric	DREAM	*Friends*
# of spelling errors	11	146
# of grammar errors	23	16
# of conciseness suggestions	6	2
# of vocabulary suggestions	18	3
General Performance	98.0	95.0
Readability Score	93.7	95.3

Table 7: Comparison of the quality of dialogues from DREAM and *Friends* (a TV show).

whose transcripts are commonly used for dialogue understanding (Chen and Choi, 2016; Ma et al., 2018). As shown in Table 7, there exist fewer spelling mistakes and the overall score is slightly higher than that of the dialogues in *Friends*. Based on the evaluated instances, articles and verb forms are the two most frequent grammar error categories (10 and 8, respectively, out of 23) in DREAM. Besides, the language tends to be less precise in DREAM, indicated by the number of vocabulary suggestions. For example, experts tend to use expressions such as *"really hot," "really beautiful," "very bad,"* and *"very important"* rather than more appropriate yet more advanced adjectives that might hinder reading comprehension of language learners with smaller vocabularies. According to the explanations provided by the tool, the readability scores for both data sets fall into the same category *"Your text is very simple and easy to read, likely to be understood by an average 5th-grader (age 10)."*

4 Approaches

We formally introduce the dialogue-based reading comprehension task and notations in Section 4.1. To investigate the effects of different kinds of general world knowledge and dialogue structure, we incorporate them into rule-based approaches (Section 4.2) as well as non-neural (Section 4.3) and neural (Section 4.4) machine learning approaches. We describe in detail preprocessing and training in Section 4.5.

4.1 Problem Formulation and Notations

We start with a formal definition of the dialogue-based multiple-choice reading comprehension task. An n-turn dialogue D is defined as $D = \{s_1 : t_1, s_2 : t_2, \ldots, s_n : t_n\}$, where s_i represents the speaker ID (e.g., *''M''* and *''W''*), and t_i represents the text of the i^{th} turn. Let Q denote the text of question, and $O_{1..3}$ denote the text of three answer options. The task is to choose the correct one from answer options $O_{1..3}$ associated with question Q given dialogue D. In this paper, we regard this task as a three-class classification problem, each class corresponding to an answer option.

For convenience, we define the following notations, which will be referred in the rest of this paper. Let D^s denote the turns spoken by speaker s in D. Formally, $D^s = \{s_{i_1} : t_{i_1}, s_{i_2} : t_{i_2}, \ldots, s_{i_m} : t_{i_m}\}$ where $\{i_1, i_2, \ldots, i_m\} = \{i \mid s_i = s\}$ and $i_1 < i_2 < \ldots < i_m$. In particular, $s = *$ denotes all the speakers. W^{D^s} and W^{O_i} denote the ordered set of the running words (excluding punctuation marks) in D^s and O_i, respectively. Questions designed for dialogue-based reading comprehension often focus on a particular speaker. If there is exactly one speaker mentioned in a question, we use s_Q to denote this target speaker. Otherwise, $s_Q = *$. For example, given the dialogue in Table 3, $s_Q =$*''M''* for Question 1 and 2, and $s_Q = *$ for Question 3.

4.2 Rule-Based Approaches

We first attempt to incorporate dialogue structure information into *sliding window* (SW), a rule-based approach developed by Richardson et al. (2013). This approach matches a bag of words constructed from a question Q and one of its answer option O_i with a given document, and calculates the TF-IDF style matching score for each answer option.

Let $\hat{D}^s$, $\hat{Q}$, and $\hat{O}_i$ be the unordered set of distinct words (excluding punctuation marks) in D^s, Q, and O_i, respectively. Instead of only regarding dialogue D as a non-conversational text snippet, we also pay special attention to the context that is relevant to the target speaker mentioned in the question. Therefore, given a target speaker s_Q, we propose to compute a *speaker-focused* sliding window score for each answer option O_i, by matching a bag of words constructed from Q and O_i with D^{s_Q} (i.e., turns spoken by s_Q). Given speaker s, we formally define the sliding window score sw of O_i as:

$$sw_i^s = \max_j \sum_{k=1\ldots|T_i|} \begin{cases} \text{ic}^s(W_{j+k}^{D^s}) & \text{if } W_{j+k}^{D^s} \in T_i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\text{ic}^s(w) = \log\left(1 + \frac{1}{\sum_i \mathbb{1}(W_i^{D^s} = w)}\right)$, $T_i = \hat{O}_i \cup \hat{Q}$, and $W_i^{D^s}$ denotes the i-th word in W^{D^s}.

Based on these definitions, we can regard sw_i^* as the general score defined in the original sliding window approach, and $sw_i^{s_Q}$ represents the speaker-focused sliding window score considering the target speaker s_Q.

Because the sliding window score ignores long-range dependencies, Richardson et al. (2013) introduce a distance-based variation (DSW), in which a word-distance based score is subtracted from the sliding window score to arrive at the final score. Similarly, we calculate the speaker-focused distance-based score given a (Q, O_i) pair and s_Q, by counting the distance between the occurrence of a word in Q and a word in O_i in D^{s_Q}. More formally, given speaker s and a set of stop words[3] U, the distance-based score d of O_i is defined as

$$d_i^s = \begin{cases} 1 & \text{if } |I_Q^s| = 0 \text{ or } |I_{O_i}^s| = 0 \\ \frac{\delta_i^s}{|W^{D^s}|-1} & \text{otherwise} \end{cases} \quad (2)$$

where $I_Q^s = (\hat{Q} \cap \hat{D}^s) - U$, $I_{O_i}^s = (\hat{O}_i \cap \hat{D}^s) - \hat{Q} - U$, and δ_i^s is the minimum number of words between an occurrence of a question word and an answer option word in W^{D^s}, plus one. The formal definition of δ_i^s is as follows.

$$\delta_i^s = \min_{W_j^{D^s} \in I_Q^s, W_k^{D^s} \in I_{O_i}^s} |j-k| + 1 \quad (3)$$

Based on these definitions, we can regard d_i^* as the distance-based score defined in the original sliding window approach, and $d_i^{s_Q}$ represents the speaker-focused distance-based score considering speaker s_Q. In addition, the final distance-based sliding window score of O_i (Richardson et al., 2013) can be formulated as

$$sw_i^* - d_i^* \quad (4)$$

Expression (4) only focuses on the general (or speaker-independent) information (i.e., sw_i^* and d_i^*); we can capture general and speaker-focused information (i.e., $sw_i^{s_Q}$, and $d_i^{s_Q}$) simultaneously by averaging them:

$$\frac{sw_i^{s_Q} + sw_i^*}{2} - \frac{d_i^{s_Q} + d_i^*}{2} \quad (5)$$

Since a large percentage of questions cannot be solved by word-level matching, we also attempt to incorporate general world knowledge into our rule-based method. We calculate cs_i^s, the maximum cosine similarity between O_i and consecutive words of the same length in W^{D^s}, as:

$$cs_i^s = \max_j \cos\left(\overline{W^{O_i}}, \overline{W_{j\ldots j+|W^{O_i}|-1}^{D^s}}\right) \quad (6)$$

where $\overline{x}$ is obtained by averaging the embeddings of the constituent words in x. Here we use ConceptNet embeddings (Speer et al., 2017) because they leverage the knowledge graph that focuses on general world knowledge. Following Expression (5), we capture both general and speaker-focused semantic information within a dialogue as follows.

$$\frac{cs_i^{s_Q} + cs_i^*}{2} \quad (7)$$

To make the final answer option selection, our rule-based method combines Expressions (5) and (7):

$$\arg\max_i \frac{sw_i^{s_Q} + sw_i^*}{2} - \frac{d_i^{s_Q} + d_i^*}{2} + \frac{cs_i^{s_Q} + cs_i^*}{2} \quad (8)$$

4.3 Feature-Based Classifier

To explore what features are effective for dialogue understanding, we first consider a gradient boosting decision tree (GBDT) classifier. Besides the conventional bag-of-words features, we primarily focus on features related to general world knowledge and dialogue structure.

- **Bag of words of each answer option**.
- **Features inspired by rule-based approaches:** We adopt the features introduced in Section 4.2, including speaker-independent scores (i.e., sw_i^* and d_i^*) and speaker-focused scores (i.e., $sw_i^{s_Q}$ and $d_i^{s_Q}$).
- **Matching position:** $p_{1..3}^{s_Q}$ and $p_{1..3}^*$, where p_i^s is the last position (in percentage) of a word in D^s that is also mentioned in O_i; 0 if none of the words in D^s is mentioned in O_i. We consider matching position because of our observation of the existence of concessions and negotiations in dialogues (Amgoud et al., 2007). We assume the facts or opinions expressed near the end of a dialogue tend to be more critical for us to answer a question.
- **Pointwise mutual information (PMI):** $pmi_{\text{max},1..3}^{s_Q}$, $pmi_{\text{max},1..3}^*$, $pmi_{\text{min},1..3}^{s_Q}$, $pmi_{\text{min},1..3}^*$, $pmi_{\text{avg},1..3}^{s_Q}$, and $pmi_{\text{avg},1..3}^*$, where $pmi_{f,i}^s$ is defined as

$$pmi_{f,i}^s = \frac{\sum_j \log f_k \frac{C_2(W_j^{O_i}, W_k^{D^s})}{C_1(W_j^{O_i}) C_1(W_k^{D^s})}}{|W^{O_i}|} \quad (9)$$

[3] We use the list of stop words from NLTK (Bird and Loper, 2004).

$C_1(w)$ denotes the word frequency of w in external copora (we use Reddit posts [Tan and Lee, 2015]), and $C_2(w_1, w_2)$ represents the co-occurrence frequency of word w_1 and w_2 within a distance $< K$ in external copora. We use PMI to evaluate the relatedness between the content of an answer option and the target-speaker-focused context based on co-occurrences of words in external corpora, inspired by previous studies on narrative event chains (Chambers and Jurafsky, 2008).

- **ConceptNet relations (CR):** $cr_{1..3,1..|R|}$. $R = \{r_1, r_2, \ldots\}$ is the set of ConceptNet relation types (e.g., *"CapableOf"* and *"PartOf"*). $cr_{i,j}$ is the number of relation triples (w_1, r_j, w_2) that appear in the ConceptNet (Speer et al., 2017), where w_1 represents a word in answer option O_i, w_2 represents a word in D, and the relation type $r_j \in R$. Similar to the motivation for using PMI, we use CR to capture the association between an answer option and the source dialogue based on raw co-occurrence counts in the commonsense knowledge base.

- **ConceptNet embeddings (CE):** Besides the lexical similarity based on string matching, we also calculate $cs^*_{1..3}$ and $cs^{s_Q}_{1..3}$, where cs^*_i and $cs^{s_Q}_i$ represent the maximum cosine similarity between O_i and consecutive words of the same length in D and D^{s_Q}, respectively (Expression (6) in Section 4.2). We use ConceptNet embeddings (Speer et al., 2017) because they leverage the general world knowledge graph.

4.4 End-To-End Neural Network

Our end-to-end neural model is based on a generative pre-trained language model (LM). We follow the framework of finetuned transformer LM (FTLM) (Radford et al., 2018) and make modifications for dialogue-based reading comprehension.

The training procedure of FTLM consists of two stages. The first stage is to learn a high-capacity language model on a large-scale unsupervised corpus of tokens $\mathcal{U} = \{u_1, \ldots, u_n\}$ by maximizing the following likelihood:

$$L_{LM}(\mathcal{U}) = \sum_i \log P(u_i \mid u_{i-k}, \ldots, u_{i-1}; \Theta) \quad (10)$$

where k is the context window size, and the conditional probability P is modeled by a multi-layer transformer decoder (Liu et al., 2018) with parameters Θ. In the second stage, the model is adapted to a labeled data set $\mathcal{C}$, where each instance consists of a sequence of input tokens $x^1, \ldots, x^m$ with a label y, by maximizing:

$$L(\mathcal{C}) = \sum_{x,y} \log P(y \mid x^1, \ldots, x^m) + \lambda L_{LM}(\mathcal{C}) \quad (11)$$

where $P(y \mid x^1, \ldots, x^m)$ is obtained by a linear + softmax layer over the final transformer block's activation, and λ is the weight for language model. For multiple-choice reading comprehension, the input tokens $x^1, \ldots, x^m$ come from the concatenation of a start token, dialogue, question, a delimiter token, answer option, and an end token; y indicates if the answer option is correct. We refer readers to Radford et al. (2018) for more details.

Because the original FTLM framework already leverages rich linguistic information from a large unlabeled corpus, which can be regarded as a type of tacit general world knowledge, we investigate whether additional dialogue structure can further improve this strong baseline. We propose *speaker embedding* to better capture dialogue structure. Specifically, in the original framework, given an input context $(u_{-k}, \ldots, u_{-1})$ of the transformer, the encoding of u_{-i} is $\boldsymbol{we}(u_{-i}) + \boldsymbol{pe}(i)$, where $\boldsymbol{we}(\cdot)$ is the word embedding, and $\boldsymbol{pe}(\cdot)$ is the position embedding. When adapting Θ to DREAM, we change the encoding to $\boldsymbol{we}(u_{-i}) + \boldsymbol{pe}(i) + \boldsymbol{se}(u_{-i}, s_Q)$, where the speaker embedding $\boldsymbol{se}(u_{-i}, s_Q)$ is (a) $\mathbf{0}$ if the token u_{-i} is not in the dialogue (i.e. it is either a start/end/delimiter token or a token in the question/option); (b) $\boldsymbol{e}_{target}$ if the token is spoken by s_Q; (c) $\boldsymbol{e}_{rest}$ if the token is in the dialogue but not spoken by s_Q. $\boldsymbol{e}_{target}$ and $\boldsymbol{e}_{rest}$ are trainable and initialized randomly. We show the overall framework in Figure 1.

4.5 Preprocessing and Training Details

For all the models, we conduct coreference resolution to determine speaker mentions of s_Q based on simple heuristics. Particularly, we map three most common speaker abbreviations (i.e., *"M"*; *"W"* and *"F"*) that appear in dialogues to their eight most common corresponding mentions (i.e., *"man," "boy," "he,"* and *"his"*; *"woman," "girl," "she,"* and *"her"*) in questions. We keep

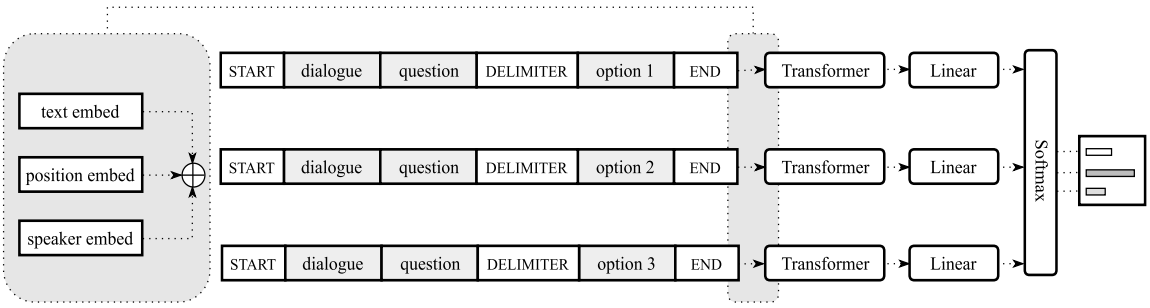

Figure 1: Overall neural network framework (Section 4.4).

speaker abbreviations unchanged, since neither replacing them with their corresponding full forms nor removing them contributes to the performance based on our experiments.

For the neural model mentioned in Section 4.4, most of our parameter settings follow Radford et al. (2018). We adopt the same preprocessing procedure and use their publicly released language model, which is pre-trained on the BooksCorpus data set (Zhu et al., 2015). We set the batch size to 8, language model weight λ to 2, and maximum epochs of training to 10.

For other models, we use the following preprocessing steps. We tokenize and lowercase the corpus, convert number words to numeric digits, normalize time expressions to 24-hour numeric form, and address negation by removing interrogative sentences that receive *''no''* as the reply. We use the gradient boosting classifier implemented in the scikit-learn toolkit (Pedregosa et al., 2011). We set the number of boosting iterations to 600 and keep the rest of hyperparameters unchanged. The distance upper bound K for PMI is set to 10.

We perform several runs of machine learning models (Section 4.3 and Section 4.4) with randomness introduced by different random seeds and/or GPU non-determinism and select the model or models (for ensemble) that perform best on the development set.

5 Experiment

5.1 Baselines

We implement several baselines, including rule-based methods and state-of-the-art neural models.

- **Word Matching** This strong baseline (Yih et al., 2013) selects the answer option that has the highest count of overlapping words with the given dialogue.

- **Sliding Window** We implement the sliding window approach (i.e., $\arg\max_i sw_i^*$) and its distance-based variation DSW (i.e., $\arg\max_i sw_i^* - d_i^*$) (Richardson et al., 2013) introduced in Section 4.2.

- **Enhanced Distance-Based Sliding Window (DSW++)** We also use general world knowledge and speaker-focused information to improve the original sliding window baseline, formulated in Expression 8 (Section 4.2).

- **Stanford Attentive Reader** This neural baseline compares each candidate answer (i.e., entity) representation to the question-aware document representation built with attention mechanism (Hermann et al., 2015; Chen et al., 2016). Lai et al. (2017) add a bilinear operation to compare document and answer option representations to answer multiple-choice questions.

- **Gated-Attention Reader** The baseline models multiplicative question-specific document representations based on a gated-attention mechanism (Dhingra et al., 2017), which are then compared to each answer option (Lai et al., 2017).

- **Co-Matching** This state-of-the-art multiple-choice reading comprehension model explicitly treats question and answer option as two sequences and jointly matches them against a given document (Wang et al., 2018b).

- **Finetuned Transformer LM** This is a general task-agnostic model introduced in Section 4.4, which achieves the best reported performance on several tasks requiring multi-sentence reasoning (Radford et al., 2018).

Method	Dev	Test
Random	32.8	33.4
Word Matching (WM) (Yih et al., 2013)	41.7	42.0
Sliding Window (SW) (Richardson et al., 2013)	42.6	42.5
Distance-Based Sliding Window (DSW) (Richardson et al., 2013)	44.4	44.6
Stanford Attentive Reader (SAR) (Chen et al., 2016)	40.2	39.8
Gated-Attention Reader (GAR) (Dhingra et al., 2017)	40.5	41.3
Co-Matching (CO) (Wang et al., 2018b)	45.6	45.5
Finetuned Transformer LM (FTLM) (Radford et al., 2018)	55.9	55.5
Our Approaches:		
DSW++ (DSW w/ Dialogue Structure and ConceptNet Embedding)	51.4	50.1
GBDT++ (GBDT w/ Features of Dialogue Structure and General World Knowledge)	53.3	52.8
FTLM++ (FTLM w/ Speaker Embedding)	**57.6**	**57.4**
Ensemble of 3 FTLM++	58.1	58.2
Ensemble of 1 GBDT++ and 3 FTLM++	**59.6**	**59.5**
Human Performance	93.9*	95.5*
Ceiling Performance	98.7*	98.6*

Table 8: Performance in accuracy (%) on the DREAM data set. Performance marked by $\star$ is reported based on 25% annotated questions from the development and test sets.

We do not investigate other ways of leveraging pre-trained deep models such as adding ELMo representations (Peters et al., 2018) as additional features to a neural model since recent studies show that directly fine-tuning a pre-trained language model such as FTLM is significantly superior on multiple-choice reading comprehension tasks (Radford et al., 2018; Chen et al., 2019). We do not apply more recent extractive models such as AOA (Cui et al., 2017) and QANet (Yu et al., 2018) since they aim at precisely locating a span in a document. When adapted to solve questions with abstractive answer options, extractive models generally tend to perform less well (Chen et al., 2016; Dhingra et al., 2017; Lai et al., 2017).

5.2 Results and Analysis

We report the performance of the baselines introduced in Section 5.1 and our proposed approaches in Table 8. We report the averaged accuracy of two annotators as the human performance. The proportion of valid questions (i.e., an unambiguous question with a unique correct answer option provided) that are manually checked by annotators on the annotated test and development sets is regarded as the human ceiling performance.

Surface matching is insufficient. Experimental results show that neural models that primarily exploit surface-level information (i.e., SAR, GAR, and CO) attain a performance level close to that of simple rule-based approaches (i.e., WM, SW, and DSW). The highest accuracy achieved by CO is 45.5%, a similar level of performance to the rule-based method DSW (44.6%).

It is helpful to incorporate general world knowledge and dialogue structure. We see a significant gain of 5.5% in accuracy when enhancing DSW using general world knowledge from ConceptNet embeddings and considering speaker-focused information (Section 4.2). FTLM, which leverages rich external linguistic knowledge from thousands of books, already achieves a much higher accuracy (55.5%) compared with previous state-of-the-art machine comprehension models, indicating the effectiveness of general world knowledge. Experimental results show that our best single model FTLM++ significantly outperforms FTLM (p-value $= 0.03$), illustrating the usefulness of additional dialogue structure. Compared with the state-of-the-art neural reader Co-Matching that primarily explores surface-level information (45.5%), the tacit general world knowledge (in the pre-trained language model) and dialogue structure in FTLM++ lead to an absolute gain of 11.9% in accuracy.

Ensembling different types of methods can bring further improvements. We use the majority vote strategy to obtain the ensemble model performance. Although GBDT++ (52.8%) itself does not outperform FTLM++, GBDT++

Method	Accuracy	Δ
DSW++	51.4	–
– dialogue structure	50.0	−1.4
– CE	46.7	−4.7
GBDT++	53.3	–
– bag of words	51.6	−1.7
– rule-based features	51.2	−2.1
– matching position	53.0	−0.3
– dialogue structure	51.9	−1.4
– PMI	51.4	−1.9
– CR	52.7	−0.6
– CE	52.7	−0.6
– PMI, CR, CE	47.1	−6.2
FTLM++	57.6	–
– speaker embedding	55.9	−1.7
– LM pre-training	36.2	−21.4

Table 9: Ablation tests on the development set (%). Minus (−) indicates percentage decrease.

can serve as a supplement to FTLM++ because they leverage different types of general world knowledge and model architectures. We achieve the highest accuracy (59.5%) by ensembling one GBDT++ and three FTLM++.

5.3 Ablation Tests

We conduct ablation tests to evaluate the individual components of our proposed approaches (Table 9). In Table 10, we summarize the involved types of dialogue structure and general world knowledge in our approaches.

Dialogue Structure Specifically, we observe 1.4% drop in accuracy if we set the target speaker s_Q to $*$ for all questions when we apply DSW++. We observe a similar performance drop when we remove speaker-focused features from GBDT++. In addition, removing speaker embeddings from FTLM++ leads to a 1.7% drop in accuracy (in this case, the model becomes the original FTLM). These results consistently indicate the usefulness of dialogue structure for dialogue understanding.

General World Knowledge We also investigate the effects of general world knowledge. The accuracy of DSW++ drops by 4.7% if we remove ConceptNet embeddings (CE) by deleting the last term of Expression (8) in Section 4.2. Additionally, the accuracy of GBDT++ drops by 6.2% if we remove all the general world knowledge features (i.e., ConceptNet embeddings/relations and PMI), leading to prediction failures on questions such as *"What do we learn about the man?"* whose correct answer option *"He is health-conscious."* is not explicitly mentioned in the source dialogue *"**M**: We had better start to eat onions frequently, Linda. **W**: But you hate onions, don't you? **M**: Until I learned from a report from today's paper that they protect people from flu and colds. After all, compared with health, taste is not so important."* Moreover, if we train FTLM++ with randomly initialized transformer weights instead of weights pre-trained on the external corpus, the accuracy drops dramatically to 36.2%, which is only slightly better than a random baseline.

	Dialogue Structure	General World Knowledge
DSW++	speaker-focused scores	CE
GBDT++	speaker-focused features	CE, CR, and PMI
FTLM++	speaker embedding	pre-trained LM

Table 10: Types of dialogue structure and general world knowledge investigated in our approaches.

5.4 Error Analysis

Impact of Longer Turns The number of dialogue turns has a significant impact on the performance of FTLM++. As shown in Figure 2, its performance reaches the peak when the number of turns ranges from 0 to 10, while it suffers severe performance drops when the given dialogue contains more turns. Both DSW++ (56.8%) and GBDT++ (57.4%) outperform FTLM++ (55.7%) when the number of turns ranges from 10 to 48. To deal with lengthy context, it may be helpful to first identify relevant sentences based on a question and its associated answer options rather than using the entire dialogue context as input.

Impact of Confusing Distractors For 54.5% of questions on the development set, the fuzzy matching score (Sikes, 2007) of at least one distractor answer option against the dialogue is higher than the score of the correct answer option. For questions that all models (i.e., DSW++, GBDT++, and FTLM++) fail to answer correctly, 73.0% of them contain at least one such confusing distractor answer option. The causes of this kind of errors can be roughly divided into two categories. First,

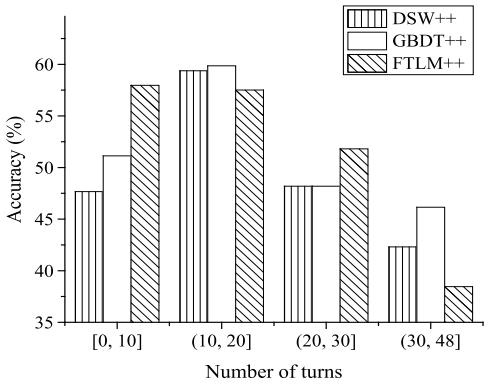

Figure 2: Performance comparison of different number of turns on the test set.

Question Type	FTLM++	GBDT++
Matching	57.0	**68.1**
Reasoning	**56.8**	49.4
Summary	**73.6**	47.1
Logic	**55.0**	49.7
Arithmetic	**30.2**	24.5
Commonsense	**53.4**	41.7
Single sentence	56.5	**63.3**
Multiple sentences	**56.9**	49.5

Table 11: Accuracy (%) by question type on the annotated development subset.

the distractor is wrongly associated with the target speaker/s mentioned in the question (e.g., answer option A and C in D2-Q3 in Table 3). Second, although the claim in the distractor is supported by the dialogue, it is irrelevant to the question (e.g., D1-Q1-B in Table 1). A promising direction to solve this problem could be the construction of speaker-focused event chains (Chambers and Jurafsky, 2008) and advanced dialogue-specific coreference resolution systems for more reliable evidence context detection in a dialogue.

Impact of Question Types We further report the performance of the best single model FTLM++ and the GBDT++ baseline on the categories defined in Section 3.2 (Table 11). Not surprisingly, both models perform worse than random guessing on math problems. While most of the math problems can be solved by one single linear equation, it is still difficult to apply recent neural math word problem solvers (Huang et al., 2018; Wang et al., 2018a) due to informal dialogue-based problem descriptions and the requirement of commonsense inference. For example, given the dialogue:
*"**W**: The plane arrives at 10:50. It is already 10:40 now. Be quick! **M**: Relax. Your watch must be fast. There are still twenty minutes left."*
We need prior knowledge to infer that the watch of the man is showing incorrect time 10:40. Instead, 10:50 should be used as the reference time with the time interval *"twenty minutes left"* together to answer the question *"What time is it now?"*

Results show that GBDT++ is superior to the fine-tuned language model on the questions under the category *matching* (68.1% vs. 57.0%) and the latter model is more capable of answering implicit questions (e.g., under the category *summary*, *logic*, and *commonsense*) which require aggregation of information from multiple sentences, the understanding of the entire dialogue, or the utilization of world knowledge. Therefore, it might be useful to leverage the strengths of individual models to solve different types of questions.

6 Conclusion and Future Work

We present DREAM, the first multiple-choice dialogue-based reading comprehension data set from English language examinations. Besides the multi-turn multi-party dialogue context, 85% of questions require multiple-sentence reasoning, and 34% of questions also require commonsense knowledge, making this task very challenging. We apply several popular reading comprehension models and find that surface-level information is insufficient. We incorporate general world knowledge and dialogue structure into rule-based and machine learning methods and show the effectiveness of these factors, suggesting a promising direction for dialogue-based reading comprehension. For future work, we are interested in problem generation for dialogues and investigating whether it will lead to more gains to pre-train a deep language model such as FTLM over large-scale dialogues from movies and TV shows instead of the BookCorpus data set (Zhu et al., 2015) used by previous work (Radford et al., 2018).

Acknowledgments

We would like to thank the editors and anonymous reviewers for their helpful feedback. We also thank Hai Wang from Toyota Technological Institute at Chicago for useful discussions and valuable comments.

References

Leila Amgoud, Yannis Dimopoulos, and Pavlos Moraitis. 2007. A unified and general framework for argumentation-based negotiation. In *Proceedings of the AAMAS*, pages 1–8. New York, NY, USA.

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2016. Embracing data abundance: Booktest data set for reading comprehension. *CoRR*, cs.CL/1610.00956v1.

Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In *Proceedings of the ACL on Interactive poster and demonstration sessions*, pages 31–34. Barcelona, Spain.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of the ACL*, pages 789–797. Columbus, OH.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the ACL*, pages 2358–2367. Berlin, Germany.

Yu-Hsin Chen and Jinho D. Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In *Proceedings of the SIGDial*, pages 90–100. Los Angeles, CA.

Zhipeng Chen, Yiming Cui, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. Convolutional spatial attention model for reading comprehension with multiple-choice questions. In *Proceedings of the AAAI*. Honolulu, HI.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. *QuAC: Question answering in context*. pages 2174–2184. Brussels, Belgium.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D. Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of the AAAI*, pages 2580–2586. Phoenix, AZ.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-attention neural networks for reading comprehension. In *Proceedings of the ACL*, pages 593–602. Vancouver, Canada.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of the ACL*, pages 1832–1846. Vancouver, Canada.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A new Q&A data set augmented with context from a search engine. *CoRR*, cs.CL/1704.05179v3.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the ACL*, pages 266–276. Vancouver, Canada.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the NIPS*, pages 1693–1701. Montreal, Canada.

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.

Danqing Huang, Jing Liu, Chin-Yew Lin, and Jian Yin. 2018. Neural math word problem solver with reinforcement learning. In *Proceedings of the COLING*, pages 213–223. Santa Fe, NM.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge data set for reading comprehension. *CoRR*, cs.CL/1705.03551v2.

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.

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 data set from examinations. In *Proceedings of the EMNLP*, pages 785–794. Copenhagen, Denmark.

Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee. 2018. ODSQA: Open-domain spoken question answering data set. *CoRR*, cs.CL/1808.02280v1.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension. *CoRR*, cs.CL/1804.00320v1.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by summarizing long sequences. In *Proceedings of the ICLR*. Vancouver, Canada.

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the ACL*, pages 329–334. Portland, OR.

Kaixin Ma, Tomasz Jurczyk, and Jinho D. Choi. 2018. Challenging reading comprehension on daily conversation: Passage completion on multiparty dialog. In *Proceedings of the NAACL-HLT*, pages 2039–2048. New Orleans, LA.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new data set for open book question answering. In *Proceedings of the EMNLP*. Brussels, Belgium.

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

I Nation. 2006. How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63:59–82.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension data set. *CoRR*, cs.CL/1611.09268v2.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did What: A large-scale person-centered cloze data set. In *Proceedings of the EMNLP*, pages 2230–2235. Austin, TX.

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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Eduard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Anselmo Penas, Yusuke Miyao, Alvaro Rodrigo, Eduard H Hovy, and Noriko Kando. 2014. Overview of CLEF QA Entrance Exams Task 2014. In *Proceedings of the CLEF*, pages 1194–1200. Sheffield, UK.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the NAACL-HLT*, pages 2227–2237, New Orleans, LA.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Preprint*, available at https://openai.com/blog/language-unsupervised/.

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. 2018. CoQA: A conversational question answering challenge. *CoRR*, cs.CL/1808.07042v1.

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

Hideyuki Shibuki, Kotaro Sakamoto, Yoshinobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly Y Itakura, Di Wang, Tatsunori Mori, and Noriko Kando. 2014. Overview of the NTCIR-11 QA-Lab Task. In *NTCIR*.

Richard Sikes. 2007. Fuzzy matching in theory and practice. *Multilingual*, 18(6):39–43.

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.

Chenhao Tan and Lillian Lee. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the WWW*, pages 1056–1066. Florence, Italy.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Multi-range reasoning for machine comprehension. *CoRR*, cs.CL/1803.09074v1.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension data set. In *Proceedings of the RepL4NLP*, pages 191–200. Vancouver, Canada.

Bo-Hsiang Tseng, Sheng-Syun Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Towards machine comprehension of spoken content: Initial toefl listening comprehension test by machine. In *Proceedings of the Interspeech*. San Francisco, CA.

Lei Wang, Yan Wang, Deng Cai, Dongxiang Zhang, and Xiaojiang Liu. 2018a. Translating a math word problem to a expression tree. In *Proceedings of the EMNLP*, pages 1064–1069. 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.

Wen-tau Yih, Ming-Wei Chang, Christopher Meek, and Andrzej Pastusiak. 2013. Question answering using enhanced lexical semantic models. In *Proceedings of the ACL*, pages 1744–1753. Sofia, Bulgaria.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining local convolution with global self-attention for reading comprehension. In *Proceedings of the ICLR*. Vancouver, Canada.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE ICCV*, pages 19–27. Santiago, Chile.