FriendsQA: Open-Domain Question Answering on TV Show Transcripts

Zhengzhe Yang
Computer Science
Emory University
Atlanta, GA, USA
zhengzhe.yang@emory.edu

Jinho D. Choi
Computer Science
Emory University
Atlanta, GA, USA
jinho.choi@emory.edu

Abstract

This paper presents FriendsQA, a challenging question answering dataset that contains 1,222 dialogues and 10,610 open-domain questions, to tackle machine comprehension on everyday conversations. Each dialogue, involving multiple speakers, is annotated with several types of questions regarding the dialogue contexts, and the answers are annotated with certain spans in the dialogue. A series of crowdsourcing tasks are conducted to ensure good annotation quality, resulting a high inter-annotator agreement of 81.82%. A comprehensive annotation analytics is provided for a deeper understanding in this dataset. Three state-of-the-art QA systems are experimented, R-Net, QANet, and BERT, and evaluated on this dataset. BERT in particular depicts promising results, an accuracy of 74.2% for answer utterance selection and an F1-score of 64.2% for answer span selection, suggesting that the FriendsQA task is hard yet has a great potential of elevating QA research on multiparty dialogue to another level.

1 Introduction

Question answering (QA) has received lots of hype over the recent years as deep learning models have progressively pushed the limit of machine comprehension to the level of human intelligence. Several systems have demonstrated their superiority over human for answering quizbowl questions (Ferrucci, 2011; Yamada et al., 2017). Strong evidences have been found that advance neural network models will likely surpass human performance for answering open-domain questions in a foreseeable future (Devlin et al., 2018; Liu et al., 2019). Nonetheless, no system has reached such high intelligence for understanding contexts in dialogue, although it is the most natural means of human communication. Moreover, the amount of data in this form has increased at a faster rate than any other type of textual data (Newport, 2014; Gonçalves, 2017).

Many datasets have been presented for various QA tasks (Section 2.1). While numerous models have shown remarkable results with these datasets (Section 2.2), the evidence passages, where the contexts of questions are derived from, mostly reside within wiki articles, newswire, (non-)fictional stories, or children's books, but not from multiparty dialogue. Contextual understanding in dialogue is challenging because it needs to interpret contents composed by multiple speakers, and anticipate colloquial language filled with sarcasms, metaphors, humors, etc. This inspires us to create a new dataset, FriendsQA, that aims to enhance machine comprehension on this domain. Dialogues in this dataset are excerpted from transcripts of the TV show *Friends*, that is the world-wide and also go-to show for English learners to get familiarized with everyday conversations.

Section 3 describes the FriendsQA dataset with annotation details. Section 4 describes the architectures of QA systems experimented on this dataset. Finally, Section 5 shows the experimental results with an in-depth error analysis. To the best of our knowledge, FriendsQA is the first dataset that is publicly available and challenges span-based QA on multiparty dialogue with everyday topics. The contributions of this work include:

- An open-domain question answering dataset on multiparty dialogue comprising 1,222 dialogues, 10,610 questions, and 21,262 answer spans.
- A comprehensive corpus analytics to ensure its validity as a deep learning resource and explain the diverse nature of this dataset for QA.
- Model comparisons between three state-of-the-art QA systems trained on this dataset to project its practicality in real applications.
- A thorough error analysis to illustrate major challenges found in this task and make suggestions to future research on the dialogue domain.

Proceedings of the SIGDial 2019 Conference, pages 188–197
Stockholm, Sweden, 11-13 September 2019.

2 Related Work

2.1 QA Datasets

The NLP community has been dedicated to produce three types of question answering (QA) datasets. The first is for reading comprehension QA, where the model picks answers for multiple choice questions regarding the evidence passages. MCTest is an open-domain dataset comprising short fictional stories (Richardson et al., 2013). RACE is a large dataset compiled from English assessments for 12-18 years old students (Lai et al., 2017). TQA gives passages from middle school science lessons and textbooks (Kembhavi et al., 2017). SciQ gives passages from science exams collected via crowdsourcing (Welbl et al., 2017). DREAM gives multiparty dialogue passages from English-as-a-foreign-language exams (Sun et al., 2019).

The second is for cloze-style QA, for which the model fills in the blanks that obliterate certain contents in sentences describing the evidence passages. CNN/Daily Mail targets on entities in bullet points summarizing articles from *CNN* and *Daily News* (Hermann et al., 2015). Children's Book Test focuses on named entities, nouns, verbs, and prepositions in passages from children's books (Hill et al., 2016). Who-did-What gives description sentences and evidence passages extracted from news articles in English Gigaword (Onishi et al., 2016). Book-Test is similar to Children's Book Test but 60 times larger (Bajgar et al., 2016).

The third is for span-based QA, where the model finds the answer contents as spans in the evidence passages. bAbI aims to reinforce learning on event types and infer a sequence of event descriptions (Weston et al., 2016). WikiQA (Yang et al., 2015) and SQuAD (Rajpurkar et al., 2016) use Wikipedia, whereas NewsQA (Trischler et al., 2017) use CNN articles as evidence passages. MS MARCO gives questions involving zero to multiple answer contents from web documents (Nguyen et al., 2016). TriviaQA is compiled by trivia enthusiasts to challenge machine comprehension (Joshi et al., 2017). CoQA focuses on conversational flows between a questioner and an answerer (Reddy et al., 2018).

2.2 QA Systems for the Past Two Years

Wang et al. (2017) presented R-Net that used gated attention-based recurrent networks and refined QA representation with self-matching attention. Shen et al. (2017) presented ReasoNet that took multiple turns to reason over the relationships between query, documents, and answers. Cui et al. (2017) presented the Attention Over Attention Reader to better capture similarities between questions and answer contents. Hu et al. (2017) presented the Reinforced Mnemonic Reader to combine the memorized attention with new attention. Vaswani et al. (2017) applied self-attention to QA, which became known as the Transformer.

Huang et al. (2018) presented FusionNet that kept the history of word representations and used multi-level attention. Salant and Berant (2018) presented a standard neural architecture with rich contextualized word representations. Liu et al. (2018) presented Stochastic Answer Network (SAN) with a stochastic prediction dropout layer as the final layer. Yu et al. (2018) presented QANet with CNN and self-attention to combine local and global interactions. Peters et al. (2018) presented the Embeddings from Language Models (ELMo) that used bi-directional LSTM and Devlin et al. (2018) presented the Bidirectional Encoder Representations (BERT) that used deep-layered transformers to generate contextualized word embeddings.

2.3 Character Mining

The Character Mining dataset provides transcripts of the TV show *Friends* as well as annotation for several tasks. Chen and Choi (2016) annotated the first two seasons for character identification, that is an entity linking task identifying personal mentions with character names. Chen et al. (2017) extended this annotation to the next two seasons and added annotation of ambiguous mentions. Zhou and Choi (2018) added annotation of plural mentions to those four seasons for character identification. Zahiri and Choi (2018) annotated the first four seasons for fine-grained emotion detection. Finally, Ma et al. (2018) annotated selected dialogues from all ten seasons for a cloze-style reading comprehension task.

2.4 FriendsQA vs. Other Dialogue QA

Three datasets have been presented for QA on dialogue. CoQA (Reddy et al., 2018) aims to answer questions that are part of one-to-one conversations, whereas FriendsQA focuses on questions asked by third-parties listening to multiparty dialogues. Ma et al. (2018) also provides a dataset based on transcripts of *Friends*; however, their work aims to cloze-style QA restricted by `PERSON` entities, while we broadly focus on span-based QA with open-domain questions. Similarly, DREAM (Sun et al., 2019), although their passages are based on

(a) Challenges with entity resolution. In this example (season 4, episode 12), $\{you_1, boys_2, us_3\}$ refer to the boys and $\{you_4, we_8\}$ refer to the girls. Many pronouns are used to refer different people, which makes it difficult to find the answer span for a question like "*who forced Rachel to raise the stakes*" by simply matching strings.

Rachel	Y'know what, **you**$_1$ are mean **boys**$_2$, who are just being mean!
Joey	Hey, don't get mad at **us**$_3$! No one forced **you**$_4$ to raise the stakes!
Rachel	That is not true. **She**$_5$ did! **She**$_6$ forced **me**$_7$!
Monica	Hey, **we**$_8$ would still be living here if **you**$_9$ hadnt gotten the question wrong!

(b) Challenges with metaphors. In this example (season 1, episode 4), Joey mishears '*omnipotent*' as "*I'm impotent*" so that he metaphorically refers it to as "*Little Joey's dead*", which makes it difficult to answer a question like "*why would Joey want to kill himself for being omnipotent*".

Monica	Hey, Joey, what would you do if you were **omnipotent**?
Joey	Probably kill myself!
Monica	Excuse me?
Joey	Hey, if **Little Joey's dead**, then I got no reason to live!

(c) Challenges with sarcasm. In this example (season 3, episode 1), Chandler is being sarcastic about him making pancakes, which makes it difficult to answer a question like "*did Chandler make pancakes*".

Chandler	Morning.
Joey	Morning, hey, you made pancakes?
Chandler	**Yeah, like there's any way I could ever do that.**

Table 1: Challenges with entity resolution, metaphors, and sarcasm in understanding dialogue contexts for QA.

dialogue, tackles multiple-choice questions, which suit well for evaluating reading comprehension, but not necessarily for practical QA applications.

3 FriendsQA Dataset

For the generation of the FriendsQA dataset, 1,222 scenes from the first four seasons of the Character Mining dataset are selected (Section 2.3). Scenes with fewer than five utterances are discarded (83 of them), and each scene is considered an independent dialogue. FriendQA can be viewed as answer span selection, where questions are asked for some contexts in a dialogue and the model is expected to find certain spans in the dialogue containing answer contents. The dialogue aspects of this dataset, however, make it more challenging than other datasets comprising passages in formal languages (Section 2.1). Three challenging aspects that are commonly found in dialogue QA are illustrated in Table 1.

3.1 Crowdsourcing

All annotation tasks are conducted on the Amazon Mechanical Turk. TALEN, a web-based tool for named entity annotation (Mayhew and Roth, 2018), is extended for our QA annotation such that it displays a dialogue segmented into a sequence of utterances with speaker names, and asks crowd workers to first generate questions then select spans or utterance IDs in the dialogue containing the answer contents (Section 3.2). Prior to the annotation, crowd workers are required to pass a quiz regarding the dialogue context, to verify if they have a good understanding in this context. Upon the submission, it validates the annotation by running several quality assurance tests (Section 3.3).

3.2 Phase 1: Question-Answer Generation

For each dialogue, the crowd workers are required to generate at least 4 out of six types of questions, $\{who, what, when, where, why, how\}$, regarding the dialogue contexts. Every question must be answerable; in other words, there needs to be at least one contiguous answer span in the dialogue. The crowd workers are allowed to select more than one answer span per question if appropriate. If multiple mentions of the same entity are to be considered, annotators are instructed to select ones that fit the best for the question. For `Q2` in Table 2, although multiple mentions of *Casey* are found in this dialogue, only the first three are selected as the answer because the other mentions are not relevant to this particular question (e.g., *Casey* in `U08`). This type of selective answer spans adds another level of difficulty to the task of FriendsQA.

Annotators are also allowed to select the speaker names as the answer spans. This is useful for *who* questions asking about certain speakers yet no mentions of them are found in the dialogue (e.g., *Chandler* has no explicit mention in Table 2). Moreover, when an entire utterance is considered the answer, which happens often with *why* and *how* questions, annotators are asked to select the corresponding utterance ID instead of the whole utterance to reduce span-related errors (e.g., `U13` for `Q5` in Table 2).

(a) A dialogue excerpted from *Friends* (season 4, episode 7).

U01	[Scene: Central Perk, Joey is getting a phone number from a woman (Casey) as Chandler watches from the doorway.]
U02	Casey: Here you go.
U03	Joey: Great! All right, so I'll call you later.
U04	Casey: Great!
U05	Chandler: Hey-Hey-Hey! Who was that?
U06	Joey: That would be Casey. We're going out tonight.
U07	Chandler: Goin' out, huh? Wow! Wow! So things didn't work out with Kathy, huh? Bummer.
U08	Joey: No, things are fine with Kathy. I'm having a late dinner with her tonight, right after my early dinner with Casey.
U09	Chandler: What?
U10	Joey: Yeah-yeah. And the craziest thing is that I just ate a whole pizza by myself!
U11	Chandler: Wait! You're going out with Kathy!
U12	Joey: Yeah. Why are you getting so upset?
U13	Chandler: Well, I'm upset for you. I mean, dating an endless line of beautiful women must be very unfulfilling for you.

(b) Six types of questions:{*who*, *what*, *when*, *where*, *why*, *how*}.

Q1	What is Joey going to do with Casey tonight?	Q4	Where are Joey and Chandler?
Q2	Who is Joey getting a phone number from?	Q5	**Why** is Chandler upset?
Q3	When will Joey have dinner with Kathy?	Q6	How are things between Joey and Kathy?

Table 2: A sample dialogue from the FriendsQA dataset comprising six types of questions, where the answer spans are annotated on the dialogue contents. Each utterance has the utterance ID, the speaker name, and the text. The answer spans for Q[1-6] are indicated by solid underlines, wavy underlines, double underlines, dashed underlines, **bold font**, and dotted underlines, respectively.

3.3 Quality Assurance

Each MTurk annotation job gives up to 6 questions and their answer spans, which are validated by the following tests before the submission:

1. Are there at least 4 types of questions annotated?
2. Does each question have at least one answer span associated with it?
3. Does any question have too much string overlaps with the original text in the dialogue?

The first test ensures that there are sufficiently large and diverse enough questions generated for developing practical QA models. The second test checks if there are any inappropriate associations between questions and answer spans. Finally, the third test prevents from creating mundane questions by copying and pasting the original text from the dialogue. No annotation job is accepted unless it passes all of these assurance tests.

3.4 Phase 2: Verification and Paraphrasing

All dialogues with the questions and answer spans annotated by the first phase (Section 3.2) are again put to the second phase. During the second phase, annotators are asked to first verify whether or not the answer spans are appropriate for the questions, and fix ones that are not or add more if necessary. Annotators are then asked to revise questions that are either unanswerable or too ambiguous. Finally, they are asked to paraphrase the questions, resulting two sets of questions for every dialogue where one is a paraphrase of the other. The same quality assurance tests (Section 3.3) with an additional test of checking string overlaps between the questions from phases 1 and 2 are run to preserve the challenging level of this dataset.

3.5 Four Rounds of Annotation

The same F1-score metric used for the evaluation of span-based QA systems (Rajpurkar et al., 2016) is used to measure the inter-annotation agreement (ITA) between the answer spans annotated by the phases 1 and 2 (Sections 3.2 and 3.4, respectively). Four rounds of crowdsourcing tasks are conducted to stabilize the quality of our annotation, where two randomly selected episodes from Seasons 1-4 are used for annotation, respectively. After each round, ITA is measured and a sample set of annotation is manually checked. Then, the annotation guidelines are updated based on this assessment. The column A from the rows R1 $\sim$ R4 in Table 3 illustrates the progressive ITA improvements over these four rounds. The followings show summaries of actions performed after each round (R[1-4]: round 1-4):

R1 We observe that the questions are often too ambiguous for humans to answer; thus, we update the guidelines and request annotators to make the questions as explicit as possible.

	S	Q	Q_p	Q_r	A	A_p	F1	$F1_p$	EM	EM_p
R1	24	122	98	62	264	216	66.59	83.42	48.15	61.17
R2	26	242	185	57	484	368	72.86	83.99	50.00	57.69
R3	30	264	213	66	528	422	75.34	83.12	48.92	53.97
R4	37	370	296	75	740	593	76.01	88.17	52.25	60.78
S1	288	2,908	2,560	627	5,824	5,123	69.93	79.78	42.78	49.01
S2	259	2,682	2,314	587	5,372	4,633	69.21	80.86	44.01	51.73
S3	291	2,908	2,546	610	5,826	5,099	72.12	81.92	47.22	53.88
S4	267	2,768	2,398	594	5,553	4,808	72.26	83.27	49.52	57.41
Total	**1,222**	12,264	**10,610**	2,678	2,4591	**21,262**	71.17	**81.82**	46.35	**53.55**

Table 3: Statistics of the FriendsQA dataset. The R[1-4] rows show the statistics for the rounds 1-4, and the S[1-4] rows show the statistics for Seasons 1-4, respectively. S: # of dialogues, Q: # of questions, Q_p: Q after pruning, Q_r: # of revised questions during phase 2, A: # of answer spans, A_p: A after pruning, F1: F1-score to measure ITA, $F1_p$: F1 after pruning, EM: exact matching score to measure ITA, EM_p: EM after pruning.

R2 We observe the 6.27% improvement on ITA from the first round; thus, we add more examples of questions and answer spans to the guidelines without updating other contents.

R3 We observe another 2.48% improvment on ITA from the second round; no update is made to the guidelines.

R4 We observe a marginal ITA improvement of 0.67% from the third round, which implies that our annotation guidelines are stabilized. Thus, all of the rest episodes are pushed for annotation.

3.6 Question/Answer Pruning

Once all annotation is collected, each question from phase 1 is represented by the bag-of-words model using TF-IDF scores and compared against its revised counterpart from phase 2 if available. About 21.8% of the questions from phase 1 are revised during phase 2. If the cosine similarity between the two questions is below 0.8, they are not considered similar so that the question and its answer spans from phase 1 are discarded because that question requires a major revision to be answerable. Even when the questions are considered similar, if the F1 score between their answer spans is below 20, they are still discarded because annotators do not seem to agree on the answer. As a result, 13.5% of the questions and answer spans from phase 1 are pruned out from our final dataset.

3.7 Inter-annotator Agreement

Table 3 show the overall statistics of the FriendsQA dataset. There is a total of 1,222 dialogues, 10,610 questions, and 21,262 answer spans in this dataset after pruning (Section 3.6). Note that annotators were not asked to paraphrase questions during the second phase of the first round (R1 in Table 3), so the number of questions in R1 is about twice less than ones from the other rounds. The final inter-annotator agreement scores are 81.82% and 53.55% for the F1 and exact matching scores respectively, indicating high-quality annotation in our dataset.

3.8 Question Types vs. Answer Categories

Table 4 shows the statistics between the question types and answer categories, where answers to each question type are further analyzed into categories. Questions show balanced distributions across different types, implying good diversity of the dataset. The analysis of answer categorization is performed manually among 250 randomly sampled questions.

Type	Count	Answer Categories (%)			
What	2,058	Factual:	100.00		
Where	1,896	Factual:	77.78	Abstract:	22.22
Who	1,847	Speaker:	30.56	Content:	69.44
Why	1,688	Explicit:	73.53	Implicit:	26.47
How	1,628	Explicit:	77.42	Implicit:	22.58
When	1,493	Absolute:	62.07	Relative:	37.93

Table 4: Statistics of the question types as well as the answer categories.

What No distinct categorization is found for answers to *what* questions, which are entirely factual. This is because annotators are mostly driven by factoid contents for the generation of *what* questions.

Where Answers to *where* questions can be categorized into factual and abstract, meaning that they are either concrete facts (e.g., named entities) or abstract concepts (e.g., *the wild*, *out there*), where the majority is driven by factoid contents (77.78%).

Who Answers to *who* questions can be annotated on either speaker names or utterance contents. The majority of *who* questions (69.44%) finds their answers in the utterance contents.

Why and How Answers to *why* and *how* questions are categorized into explicit and implicit such that they are either directly answering the questions (e.g., why doesn't Joey want to throw the chair out? → *Joey: I built this thing with my own hand*), or indirectly implying the answers (e.g., How are Joey and Chandler going to get to Monica's place? → *Joey: we're not gonna have to walk there, right?*). Explicit answers are more common for both *why* (73.53%) and *how* (77.42%) questions.

When Answers to *when* questions can be categorized into absolute and relative such that they can be either exact timing (e.g., clock time, specific date, holiday) or timing of action relative to another event (e.g., I called her *while I was watching TV*). About two third of the answers are considered explicit for *when* questions.

4 State-of-the-Art QA Systems

Three of state-of-the-art QA systems, R-Net based on recurrent neural networks (RNN) (Section 4.1), QANet based on convolutional neural networks (CNN) with self-attention (Section 4.2), and BERT based on deep feed-forward neural networks with transformers (Section 4.3), are used to validate our dataset as a practical resource for building advanced deep learning models. All models will output two positions which will be combined to form answer spans. These systems are chosen because they give a good survey among different types of neural networks in combination with attention mechanisms that are dominant in the research of contemporary question answering.

4.1 R-Net

R-Net held the 1st place on the SQuAD leaderboard at the time of its publication (Wang et al., 2017). It builds representations for questions and evidence passages using RNN and presents a self-matching mechanism to aggregate key information from the evidence passages, in order to compensate the limitedly memorized information from RNN. The same configuration described in the original paper is used to train models for our experiments.

4.2 QANet

QANet is another state-of-the-art open-domain QA system utilizing CNN and self-attention (Yu et al., 2018). Dramatic is the speed-up gained by QANet, which enables to perform data augmentation. Their original configuration cannot be fit in a 12GB GPU machine using our dataset; thus, the configuration is compromised for our experiments as follows:

- The number of filters: 96 instead of 128,
- The number of attention heads: 1 instead of 8.

Given this configuration, its performance may not be optimal but at least can be directly compared to other models trained on the FriendsQA dataset.

4.3 BERT

The Bidirectional Encoder Representations from Transformers (BERT) pushed all current state-of-the-art scores to another level (Devlin et al., 2018). Trained with the masked language model on next sentence prediction tasks, BERT shows extremely promising results on several tasks in NLP. The pre-trained decapitalized BERT model with 12-layers is fine-tuned on our dataset. The larger BERT model with 24-layers again cannot be fit in a 12GB GPU machine; thus, it is not used for our experiments.

5 Experiments

For our experiments, all dialogues from Table 3 are randomly shuffled and redistributed as the training (80%), development (10%), and test (10%) as shown in Table 5.

Set	Dialogues	Questions	Answers
Training	977	8,535	17,074
Development	122	1,010	2,057
Test	123	1,065	2,131

Table 5: Data split for our experiments.

5.1 Model Development

Each instance consists of an evidence dialogue, a question and an answer span. Utterance IDs, annotated to indicate the whole utterances being answer spans (Section 3.2), are preprocessed and replaced by the actual spans on the dialogue contents. Since each question can have multiple answers, the following strategies are experimented to acquire one gold answer span for each training instance:

Shortest The shortest answer span is chosen and all the other spans are discarded from training.

Longest The longest answer span is chosen and all the other spans are discarded from training.

Multiple The question is paired with every answer to create multiple instances. For example, a question q with two answer spans, a_1 and a_2, generate two instances, (q, a_1) and (q, a_2), and trained independently.

Model	Shortest-Answer Strategy			Longest-Answer Strategy			Multiple-Answer Strategy		
	UM	SM	EM	UM	SM	EM	UM	SM	EM
R-Net	45.41 (±1.16)	35.69 (±1.28)	**25.55** (±1.60)	**49.50** (±0.54)	**37.26** (±0.72)	23.77 (±0.42)	43.77 (±0.56)	33.97 (±0.75)	23.02 (±1.30)
QANet	42.12 (±3.21)	34.04 (±0.03)	22.89 (±0.42)	46.21 (±4.51)	34.55 (±1.87)	21.15 (±1.21)	**47.10** (±1.30)	**35.38** (±1.33)	**23.16** (±1.15)
BERT	72.61 (±0.20)	63.64 (±0.42)	48.33 (±1.41)	72.16 (±1.93)	60.36 (±1.53)	43.23 (±1.83)	**74.18** (±0.21)	**64.15** (±0.29)	**48.96** (±0.42)

Table 6: Results from the three state-of-the-art QA systems. All models are experimented three times and their average scores with standard deviations are reported. UM: Utterance Match, SM: Span Match, EM: Exact Match.

5.2 Evaluation Metrics

Two tasks are experimented, answer utterance selection and answer span selection, with the FriendsQA dataset. The utterance match (UM) is used to evaluate answer utterance selection, which checks if the predicted answer span a_i^p resides within the same utterance u_i^g as the gold answer span a_i^g, and is measured as follows: (n: # of questions):

$$\textbf{UM} = \frac{1}{n}\sum_{i=1}^{n}(1 \text{ if } a_i^p \in u_i^g; \text{otherwise}, 0)$$

Following Rajpurkar et al. (2016), the span match (SM) is adapted to evaluate answer span selection, where each a_i^p is treated as a bag-of-tokens (ϕ) and compared to the bag-of-tokens of a_i^g; the macro-average F1 score across all questions is measured for the final evaluation (P: precision, R: recall):

$$\textbf{SM} = \frac{1}{n}\sum_{i=1}^{n}\frac{2 \cdot P(\phi(a_i^p), \phi(a_i^g))R(\phi(a_i^p), \phi(a_i^g))}{P(\phi(a_i^p), \phi(a_i^g)) + R(\phi(a_i^p), \phi(a_i^g))}$$

Additionally, the exact match (EM) is used to evaluate answer span selection that checks the exact span match between the gold and predicted answers.

5.3 Results

Table 6 shows results from 9 models trained by the three state-of-the-art systems in Section 5.2 using the three answer selection strategies in Section 5.1. All experiments are run three times and their average scores with standard deviations are reported. BERT and QANet perform better with the multiple-answer strategy, that gives more training instances per question, whereas R-Net performs better with the other strategies. This could be due to R-Net's self-matching mechanism that gets confused when multiple answers are provided for training the same question. BERT models significantly outperform ones from the other two systems in all evaluations. Since our hyper-parameters are tuned around grids provided by the original papers, it is possible that these results are still suboptimal, which points out another important property of BERT that it is not as sensitive to different QA datasets.

Type	Dist.	UM	SM	EM
What	19.70%	77.43	69.39	55.04
Where	18.28%	84.35	78.86	65.93
Who	17.17%	74.12	64.34	55.29
Why	15.76%	60.47	50.03	27.14
How	14.65%	65.52	52.04	32.64
When	14.44%	80.65	65.81	51.98

Table 7: Results with respect to question types using BERT and the multiple-answer strategy.

Table 7 shows results from BERT's multiple answer models by question types. Answers to *where* and *when* questions are mostly factoid, which show the highest performance. On the other hand, answers to *why* and *how* usually span out to longer sequences, leading to worse performance. Answers to *who* and *what* questions give a good mixture of proper and common nouns and show moderate performance.

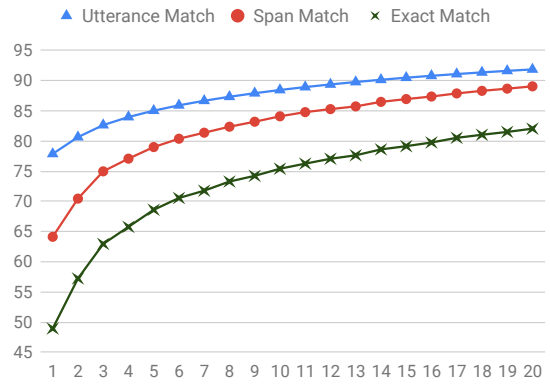

Figure 1: Increasing score with top-20 answer candidates. From top to bottom: Utterance Match, Span Match and Exact Match.

Figure 1 shows improvement of BERT's multiple-answer models by accepting the top-k answer predictions; the scores are measured by picking the best matching answer within thes top-k predictions. UM surpasses 90% and SM approaches to 90%

when k = 14 and 20, respectively. More importantly, the gap between UM and SM gets smaller as k increases, which implies that FriendsQA is not only learnable by deep learning but also can be enhanced by re-ranking the answer predictions.

5.4 Error Analysis

An extensive error analysis is manually performed on 100 randomly sampled, exact unmatched predictions (F1 = 0) to provide insights for future research. Figure 2 shows six types of errors that become evident through this analysis.

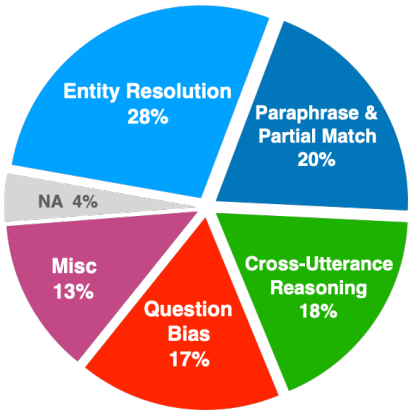

Figure 2: The distribution of six error types analyzed in 100 sampled predictions. NA: Noise in annotation.

Entity Resolution This type is the most frequent and often occurs when many of the same entities are mentioned in multiple utterances. The recurring use of coreference and anaphora can be confusing. This error also occurs when the QA system is asked about a specific person, but predicts wrong people. For example, the question asks for Chandler's opinion about marriage, but the model matches comments from Joey instead due to the lack of referent resolution made in those comments.

Paraphrase and Partial Match This error type may be even challenging for humans without inside knowldege. Answers can be expressed in numerous ways through paraphrasing, abstraction, nicknames in dialogue, signifying the difficulty in FriendsQA. Moreover, answers might also be partially correct, especially for *why* and *how* questions, which could be acceptable in practice.

Cross-Utterance Reasoning This type reveals an universal challenge in understanding human-to-human conversation. To correctly predict an answer span in dialogue, the system should be equipped with the ability to reason across multiple utterances back and forth, especially if a story or an event unfolds gradually, scatters in different places, and is told by different speakers.

Question Bias This type occurs when the answer predictions overly rely on the question types. For *why* questions, the model tends to blindly selects spans following certain keywords such as *because* even though they are placed in wrong utterances since the model is learned to be biased to the term *because*, neglecting other important factors that might otherwise lead to the correct answers.

Noise in Annotation (NA) Our dataset, although it gives high inter-annotator agreement (Sec. 3.7), it still includes noise caused by wrong spans, ambiguous or unanswerable questions, or typos.

Miscellaneous Errors in this category have no apparent cause to understand why the model predicts these answers, which often seem irrelevant to the questions so that they need more investigation.

Given this analysis, we hope many challenges can be overcome by future studies. For instance, coreferent mentions, especially plural mentions, should be more intelligently processed (Zhou and Choi, 2018). Moreover, the speaker information, which are currently treated as the first tokens in utterances, can be better encoded to give more insights.

6 Conclusion

This paper presents an open-domain question answering dataset called FriendsQA, compiled from transcripts of the TV show *Friends*. An extensive and comprehensive analysis is performed on this dataset to show its validity, difficulty and diversity. Three state-of-the-art models are run and compared, and show the full potential of FriendsQA as a rich QA research resource. Finally, erroneous answer predictions are sampled out for a further analysis to offer insightful retrospective. All our resources are publicly available.[1]

For future work, the question-type (Table 7) and error analyses (Section 5.4) can serve as guidelines to further enhance the QA model performance. Top-k answer analysis also brings up another challenging but tangible task to re-rank the answer predictions. More tasks such as answer existence prediction and an utterance-based model to select among utterance candidates can also be issued.

[1]FriendsQA: https://github.com/emorynlp/question-answering

References

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2016. Embracing data abundance: BookTest Dataset for Reading Comprehension. *arXiv*, 1610.00956.

Henry Yu-Hsin Chen and Jinho D. Choi. 2016. Character Identification on Multiparty Conversation: Identifying Mentions of Characters in TV Shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL'16, pages 90–100.

Henry Yu-Hsin Chen, Ethan Zhou, and Jinho D. Choi. 2017. Robust Coreference Resolution and Entity Linking on Dialogues: Character Identification on TV Show Transcripts. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, CoNLL'17.

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 55th Annual Meeting of the Association for Computational Linguistics*, ACL'17, pages 593–602.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, 1810.04805.

David A. Ferrucci. 2011. IBM's Watson/DeepQA. In *Proceedings of the 38th Annual International Symposium on Computer Architecture*, ISCA'11.

Pedro Gonçalves. 2017. 10 graphs that show why your business should be available through messaging apps.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Annual Conference on Neural Information Processing Systems*, NIPS'15, pages 1693–1701.

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 6th International Conference on Learning Representations*, ICLR'16.

Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2017. Reinforced Mnemonic Reader for Machine Reading Comprehension. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pages 4099–4106.

Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2018. FusionNet: Fusing via Fully-aware Attention with Application to Machine Comprehension. In *Proceedings of the International Conference on Learning Representations*, page ICLR'18.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL'17, pages 1601–1611.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are You Smarter Than a Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension. In *The IEEE Conference on Computer Vision and Pattern Recognition*, CVPR'17.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'17, pages 785–794.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-Task Deep Neural Networks for Natural Language Understanding. *arXiv*, 1901.11504.

Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2018. Stochastic Answer Networks for Machine Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL'18, pages 1694–1704.

Kaixin Ma, Tomasz Jurczyk, and Jinho D. Choi. 2018. Challenging Reading Comprehension on Daily Conversation: Passage Completion on Multiparty Dialog. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2039–2048, New Orleans, Louisiana. Association for Computational Linguistics.

Stephen Mayhew and Dan Roth. 2018. TALEN: Tool for Annotation of Low-resource ENtities. In *Proceedings of the ACL System Demonstrations*, ACL:DEMO'18, pages 80–86.

Frank Newport. 2014. The New Era of Communication Among Americans.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the NIPS Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did What: A Large-Scale Person-Centered Cloze Dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP'16, pages 2230–2235.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL'18, pages 2227–2237.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'16, pages 2383–2392.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. CoQA: A Conversational Question Answering Challenge. *arXiv*, 1808.07042.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP'13, pages 193–203.

Shimi Salant and Jonathan Berant. 2018. Contextualized Word Representations for Reading Comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL'18, pages 554–559.

Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. ReasoNet: Learning to Stop Reading in Machine Comprehension. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'17, pages 1047–1055.

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 for Computational Linguistics*, 7.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, NIPS'17, pages 5998–6008.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated Self-Matching Networks for Reading Comprehension and Question Answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL'17, pages 189–198.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing Multiple Choice Science Questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2016. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR'16.

Ikuya Yamada, Ryuji Tamaki, Hiroyuki Shindo, and Yoshiyasu Takefuji. 2017. Studio Ousia's Quiz Bowl Question Answering System at NIPS HCQA 2017. In *Human–Computer Question Answering Competition at the 31 Annual Conference on Neural Information Processing Systems*, NIPS-HCQA'17.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WIKIQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'15, pages 2013–2018.

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 6th International Conference on Learning Representations*, ICLR'18.

Sayyed Zahiri and Jinho D. Choi. 2018. Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks. In *Proceedings of the AAAI Workshop on Affective Content Analysis*, AFFCON'18, New Orleans, LA.

Ethan Zhou and Jinho D. Choi. 2018. They Exist! Introducing Plural Mentions to Coreference Resolution and Entity Linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34.