Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading

Lianhui Qin[†], **Michel Galley**[‡], **Chris Brockett**[‡], **Xiaodong Liu**[‡], **Xiang Gao**[‡], **Bill Dolan**[‡], **Yejin Choi**[†] and **Jianfeng Gao**[‡]
[†] University of Washington, Seattle, WA, USA
[‡] Microsoft Research, Redmond, WA, USA
{lianhuiq,yejin}@cs.washington.edu
{mgalley,Chris.Brockett,xiaodl,xiag,billdol,jfgao}@microsoft.com

Abstract

Although neural conversation models are effective in learning *how* to produce fluent responses, their primary challenge lies in knowing *what* to say to make the conversation *contentful* and non-vacuous. We present a new end-to-end approach to contentful neural conversation that jointly models response generation and on-demand machine reading. The key idea is to provide the conversation model with relevant long-form text *on the fly* as a source of external knowledge. The model performs QA-style reading comprehension on this text in response to each conversational turn, thereby allowing for more focused integration of external knowledge than has been possible in prior approaches. To support further research on knowledge-grounded conversation, we introduce a new large-scale conversation dataset grounded in external web pages (2.8M turns, 7.4M sentences of grounding). Both human evaluation and automated metrics show that our approach results in more contentful responses compared to a variety of previous methods, improving both the informativeness and diversity of generated output.

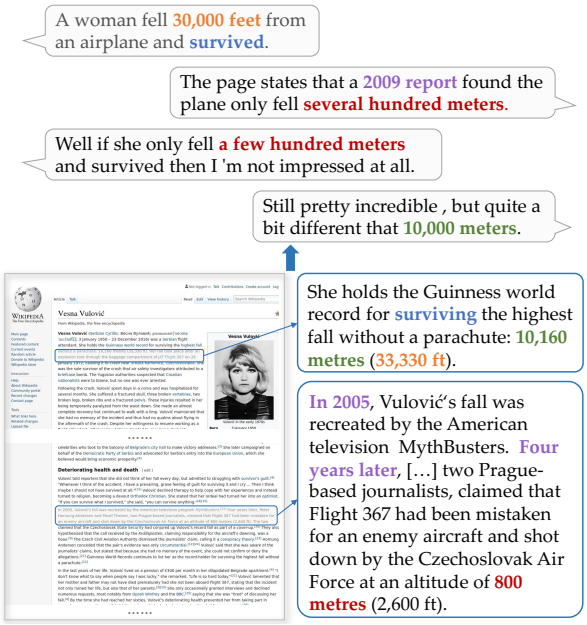

Figure 1: Users discussing a topic defined by a Wikipedia article. In this real-world example from our Reddit dataset, information needed to ground responses is distributed throughout the source document.

1 Introduction

While end-to-end neural conversation models (Shang et al., 2015; Sordoni et al., 2015; Vinyals and Le, 2015; Serban et al., 2016; Li et al., 2016a; Gao et al., 2019a, etc.) are effective in learning *how* to be fluent, their responses are often vacuous and uninformative. A primary challenge thus lies in modeling *what* to say to make the conversation contentful. Several recent approaches have attempted to address this difficulty by conditioning the language decoder on external information sources, such as knowledge bases (Agarwal et al., 2018; Liu et al., 2018a), review posts (Ghazvininejad et al., 2018; Moghe et al., 2018), and even images (Das et al., 2017; Mostafazadeh et al., 2017). However, empirical results suggest that conditioning the decoder on rich and complex contexts, while helpful, does not on its own provide sufficient inductive bias for these systems to learn how to achieve deep and accurate integration between external knowledge and response generation.

We posit that this ongoing challenge demands a more effective mechanism to support on-demand knowledge integration. We draw inspiration from how humans converse about a topic, where people often search and acquire external information as needed to continue a meaningful and informative conversation. Figure 1 illustrates an example human discussion, where information scattered in separate paragraphs must be consolidated to com-

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5427–5436
Florence, Italy, July 28 - August 2, 2019.

pose grounded and appropriate responses. Thus, the challenge is to connect the dots across different pieces of information in much the same way that *machine reading comprehension (MRC)* systems tie together multiple text segments to provide a unified and factual answer (Seo et al., 2017, etc.).

We introduce a new framework of end-to-end conversation models that jointly learn response generation together with on-demand machine reading. We formulate the reading comprehension task as document-grounded response generation: given a long document that supplements the conversation topic, along with the conversation history, we aim to produce a response that is both conversationally appropriate and informed by the content of the document. The key idea is to project conventional QA-based reading comprehension onto conversation response generation by equating the conversation prompt with the question, the conversation response with the answer, and external knowledge with the context. The MRC framing allows for integration of long external documents that present notably richer and more complex information than relatively small collections of short, independent review posts such as those that have been used in prior work (Ghazvininejad et al., 2018; Moghe et al., 2018).

We also introduce a large dataset to facilitate research on knowledge-grounded conversation (2.8M turns, 7.4M sentences of grounding) that is at least one order of magnitude larger than existing datasets (Dinan et al., 2019; Moghe et al., 2018). This dataset consists of real-world conversations extracted from Reddit, linked to web documents discussed in the conversations. Empirical results on our new dataset demonstrate that our full model improves over previous grounded response generation systems and various ungrounded baselines, suggesting that deep knowledge integration is an important research direction.[1]

2 Task

We propose to use factoid- and entity-rich web documents, e.g., news stories and Wikipedia pages, as external knowledge sources for an open-ended conversational system to ground in.

Formally, we are given a conversation history of turns $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M)$ and a web document $D = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)$ as the knowledge source, where $\boldsymbol{s}_i$ is the ith sentence in the document. With the pair (X, D), the system needs to generate a natural language response $\boldsymbol{y}$ that is both conversationally appropriate and reflective of the contents of the web document.

3 Approach

Our approach integrates conversation generation with on-demand MRC. Specifically, we use an MRC model to effectively encode the conversation history by treating it as a question in a typical QA task (e.g., SQuAD (Rajpurkar et al., 2016)), and encode the web document as the context. We then replace the output component of the MRC model (which is usually an answer classification module) with an attentional sequence generator that generates a free-form response. We refer to our approach as CMR (Conversation with on-demand Machine Reading). In general, any off-the-shelf MRC model could be applied here for knowledge comprehension. We use Stochastic Answer Networks (SAN)[2] (Liu et al., 2018b), a performant machine reading model that until very recently held state-of-the-art performance on the SQuAD benchmark. We also employ a simple but effective data weighting scheme to further encourage response grounding.

3.1 Document and Conversation Reading

We adapt the SAN model to encode both the input document and conversation history and forward the digested information to a response generator. Figure 2 depicts the overall MRC architecture. Different blocks capture different concepts of representations in both the input conversation history and web document. The leftmost blocks represent the lexicon encoding that extracts information from X and D at the token level. Each token is first transformed into its corresponding word embedding vector, and then fed into a position-wise feed-forward network (FFN) (Vaswani et al., 2017) to obtain the final token-level representation. Separate FFNs are used for the conversation history and the web document.

The next block is for contextual encoding. The aforementioned token vectors are concatenated with pre-trained 600-dimensional CoVe vectors (McCann et al., 2017), and then fed to a BiL-

[1] Code for reproducing our models and data is made publicly available at https://github.com/qkaren/converse_reading_cmr.

[2] https://github.com/kevinduh/san_mrc

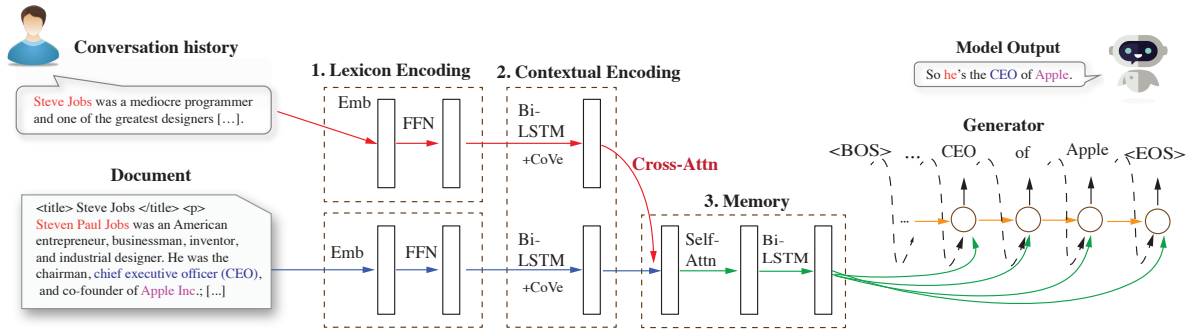

Figure 2: **Model Architecture for Response Generation with on-demand Machine Reading:** The first blocks of the MRC-based encoder serve as a lexicon encoding that maps words to their embeddings and transforms with position-wise FFN, independently for the conversation history and the document. The next block is for contextual encoding, where BiLSTMs are applied to the lexicon embeddings to model the context for both conversation history and document. The last block builds the final encoder memory, by sequentially applying cross-attention in order to integrate the two information sources, conversation history and document, self-attention for salient information retrieval, and a BiLSTM for final information rearrangement. The response generator then attends to the memory and generates a free-form response.

STM that is shared for both conversation history and web document. The step-wise outputs of the BiLSTM carry the information of the tokens as well as their left and right context.

The last block builds the memory that summarizes the salient information from both X and D. The block first applies *cross*-attention to integrate information from the conversation history X into the document representation. Each contextual vector of the document D is used to compute attention (similarity) distribution over the contextual vectors of X, which is concatenated with the weighted average vector of X by the resulting distribution. Second, a *self*-attention layer is applied to further ingest and capture the most salient information. The output memory, $M \in \mathbb{R}^{d \times n}$, is obtained by applying another BiLSTM layer for final information rearrangement. Note that d is the hidden size of the memory and n is the length of the document.

3.2 Response Generation

Having read and processed both the conversation history and the extra knowledge in the document, the model then produces a free-form response $\mathbf{y} = (y_1, \ldots, y_T)$ instead of generating a span or performing answer classification as in MRC tasks.

We use an attentional recurrent neural network decoder (Luong et al., 2015) to generate response tokens while attending to the memory. At the beginning, the initial hidden state $\boldsymbol{h}_0$ is the weighted sum of the representation of the history X. For each decoding step t with a hidden state $\boldsymbol{h}_t$, we generate a token y_t based on the distribution:

$$p(y_t) = \text{softmax}((W_1 \boldsymbol{h}_t + \boldsymbol{b})/\tau), \qquad (1)$$

where $\tau > 0$ is the softmax temperature. The hidden state $\boldsymbol{h}_t$ is defined as follows:

$$\boldsymbol{h}_t = W_2[\boldsymbol{z}_t +\!\!+ f_{attention}(\boldsymbol{z}_t, M)]. \qquad (2)$$

Here, $[\cdot +\!\!+ \cdot]$ indicates a concatenation of two vectors; $f_{attention}$ is a dot-product attention (Vaswani et al., 2017); and $\boldsymbol{z}_t$ is a state generated by $\text{GRU}(\boldsymbol{e}_{t-1}, \boldsymbol{h}_{t-1})$ with $\boldsymbol{e}_{t-1}$ being the embedding of the word y_{t-1} generated at the previous $(t-1)$ step. In practice, we use top-k sample decoding to draw y_t from the above distribution $p(y_t)$. Section 5 provides more details about the experimental configuration.

3.3 Data Weighting Scheme

We further propose a simple data weighting scheme to encourage the generation of grounded responses. The idea is to bias the model training to fit better to those training instances where the ground-truth response is more closely relevant to the document. More specifically, given a training instance $(X, D, \boldsymbol{y})$, we measure the closeness score $c \in \mathbb{R}$ between the document D and the gold response $\boldsymbol{y}$ (e.g., with the NIST (Doddington, 2002) or BLEU (Papineni et al., 2002) metrics). In each training data batch, we normalize the closeness scores of all the instances to have a sum of 1, and weight each of the instances with its corresponding normalized score when evaluating the

	Train	Valid	Test
# dialogues	28.4k	1.2k	3.1k
# utterances	2.36M	0.12M	0.34M
# documents	28.4k	1.2k	3.1k
# document sentences	15.18M	0.58M	1.68M
Average length (# words):			
utterances	18.74	18.84	18.48
document sentences	13.72	14.17	14.15

Table 1: Our grounded conversational dataset.

training loss. This training regime promotes instances with grounded responses and thus encourages the model to better encode and utilize the information in the document.

4 Dataset

To create a grounded conversational dataset, we extract conversation threads from Reddit, a popular and large-scale online platform for news and discussion. In 2015 alone, Reddit hosted more than 73M conversations.[3] On Reddit, user submissions are categorized by topics or "subreddits", and a submission typically consists of a submission title associated with a URL pointing to a news or background article, which initiates a discussion about the contents of the article. This article provides framing for the conversation, and this can naturally be seen as a form of grounding. Another factor that makes Reddit conversations particularly well-suited for our conversation-as-MRC setting is that a significant proportion of these URLs contain named anchors (i.e., '#' in the URL) that point to the relevant passages in the document. This is conceptually quite similar to MRC data (Rajpurkar et al., 2016) where typically only short passages within a larger document are relevant in answering the question.

We reduce spamming and offensive language by manually curating a list of 178 relatively "safe" subreddits and 226 web domains from which the web pages are extracted. To convert the web page of each conversation into a text document, we extracted the text of the page using an html-to-text converter,[4] while retaining important tags such as <title>, <h1> to <h6>, and <p>. This means the entire text of the original web page is preserved, but these main tags retain some high-level structure of the article. For web URLs with named anchors, we preserve that information by indicating the anchor text in the document with tags <anchor> and </anchor>. As the whole documents in the dataset tend to be lengthy, anchors offer important hints to the model about which parts of the documents should likely be focused on in order to produce a good response. We considered it sensible to keep them as they are also available to the human reader.

After filtering short or redacted turns, or which quote earlier turns, we obtained 2.8M conversation instances respectively divided into train, validation, and test (Table 1). We used different date ranges for these different sets: years 2011-2016 for train, Jan-Mar 2017 for validation, and the rest of 2017 for test. For the test set, we select conversational turns for which 6 or more responses were available, in order to create a multi-reference test set. Given other filtering criteria such as turn length, this yields a 6-reference test set of size 2208. For each instance, we set aside one of the 6 human responses to assess human performance on this task, and the remaining 5 responses serve as ground truths for evaluating different systems.[5] Table 1 provides statistics for our dataset, and Figure 1 presents an example from our dataset that also demonstrates the need to combine conversation history and background information from the document to produce an informative response.

To enable reproducibility of our experiments, we crawled web pages using Common Crawl (http://commoncrawl.org), a service that crawls web pages and makes its historical crawls available to the public. We also release the code (URL redacted for anonymity) to recreate our dataset from both a popular Reddit dump[6] and Common Crawl, and the latter service ensures that anyone reproducing our data extraction experiments would retrieve exactly the same web pages. We made a preliminary version of this dataset available for a shared task (Galley et al., 2019) at Dialog System Technology Challenges (DSTC) (Yoshino et al., 2019). Back-and-forth with participants helped us iteratively refine the dataset. The code to recreate this dataset is included.[7]

[3] https://redditblog.com/2015/12/31/reddit-in-2015/

[4] https://www.crummy.com/software/BeautifulSoup

[5] While this is already large for a grounded dataset, we could have easily created a much bigger one given how abundant Reddit data is. We focused instead on filtering out spamming and offensive language, in order to strike a good balance between data quality and size.

[6] http://files.pushshift.io/reddit/

[7] We do not report on shared task systems here, as these

5 Experiments

5.1 Systems

We evaluate our systems and several competitive baselines:

SEQ2SEQ (Sutskever et al., 2014) We use a standard LSTM SEQ2SEQ model that only exploit the conversation history for response generation, without any grounding. This is a competitive baseline initialized using pretrained embeddings.

MEMNET: We use a Memory Network designed for grounded response generation (Ghazvininejad et al., 2018). An end-to-end memory network (Sukhbaatar et al., 2015) encodes conversation history and sentences in the web documents. Responses are generated with a sequence decoder.

CMR-F : To directly measure the effect of incorporating web documents, we compare to a baseline which omits the document reading component of the full model (Figure 2). As with the SEQ2SEQ approach, the resulting model generates responses solely based on conversation history.

CMR: To measure the effect of our data weighting scheme, we compare to a system that has identical architecture to the full model, but is trained without associating weights to training instances.

CMR+W: As described in section 3, the full model reads and comprehends both the conversation history and document using an MRC component, and sequentially generates the response. The model is trained with the data weighting scheme to encourage grounded responses.

Human: To get a better sense of the systems' performance relative to an upper bound, we also evaluate human-written responses using different metrics. As described in Section 4, for each test instance, we set aside one of the 6 human references for evaluation, so the 'human' is evaluated against the other 5 references for automatic evaluation. To make these results comparable, all the systems are also automatically evaluated against the same 5 references.

systems do not represent our work and some of these systems have no corresponding publications. Along with the data described here, we provided a standard SEQ2SEQ baseline to the shared task, which we improved for the purpose of this paper (improved BLEU, NIST and METEOR). Our new SEQ2SEQ baseline is described in Section 5.

6 Experiment Details

For all the systems, we set word embedding dimension to 300 and used the pretrained GloVe[8] for initialization. We set hidden dimensions to 512 and dropout rate to 0.4. GRU cells are used for SEQ2SEQ and MEMNET (we also tested LSTM cells and obtained similar results). We used the Adam optimizer for model training, with an initial learning rate of 0.0005. Batch size was set to 32. During training, all responses were truncated to have a maximum length of 30, and maximum query length and document length were set to 30, 500, respectively. we used regular teacher-forcing decoding during training. For inference, we found that top-k random sample decoding (Fan et al., 2018) provides the best results for all the systems. That is, at each decoding step, a token was drawn from the k most likely candidates according to the distribution over the vocabulary. Similar to recent work (Fan et al., 2018; Edunov et al., 2018), we set $k = 20$ (other common k values like 10 gave similar results). We selected key hyperparameter configurations on the validation set.

6.1 Evaluation Setup

Table 2 shows automatic metrics for quantitative evaluation over three qualities of generated texts. We measure the overall **relevance** of the generated responses given the conversational history by using standard Machine Translation (MT) metrics, comparing generated outputs to ground-truth responses. These metrics include BLEU-4 (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007). and NIST (Doddington, 2002). The latter metric is a variant of BLEU that weights n-gram matches by their information gain by effectively penalizing uninformative n-grams (such as "I don't know"), which makes it a relevant metric for evaluating systems aiming diverse and informative responses. MT metrics may not be particularly adequate for our task (Liu et al., 2016), given its focus on the informativeness of responses, and for that reason we also use two other types of metrics to measure the level of grounding and diversity.

As a **diversity** metric, we count all n-grams in the system output for the test set, and measure: (1) Entropy-n as the entropy of the n-gram count distribution, a metric proposed in (Zhang et al., 2018b); (2) Distinct-n as the ratio between the

[8] https://nlp.stanford.edu/projects/glove/

	Appropriateness			Grounding			Diversity			
	NIST	BLEU	METEOR	Precision	Recall	F1	Entropy-4	Distinct-1	Distinct-2	Len
Human	2.650	3.13%	8.31%	2.89%	0.45%	0.78%	10.445	0.167	0.670	18.757
SEQ2SEQ	2.223	1.09%	7.34%	1.20%	0.05%	0.10%	9.745	0.023	0.174	15.942
MEMNET	2.185	1.10%	7.31%	1.25%	0.06%	0.12%	9.821	0.035	0.226	15.524
CMR-F	**2.260**	1.20%	7.37%	1.68%	0.08%	0.15%	9.778	0.035	0.219	15.471
CMR	2.213	**1.43%**	7.33%	2.44%	0.13%	0.25%	9.818	0.046	0.258	15.048
CMR+W	2.238	1.38%	**7.46%**	**3.39%**	**0.20%**	**0.38%**	**9.887**	**0.052**	**0.283**	15.249

Table 2: **Automatic Evaluation** results (higher is better for all metrics). Our best models (CMR+W and CMR) considerably increase the quantitative measures of Grounding, and also slightly improve Diversity. Automatic measures of Quality (e.g., BLEU-4) give mixed results, but this is reflective of the fact that we did not aim to improve response relevance with respect to the context, but instead its level of grounding. The human evaluation results in Table 3 indeed suggest that our best system (CMR+W) is better.

number of n-gram types and the total number of n-grams, a metric introduced in (Li et al., 2016a).

For the **grounding** metrics, we first compute '#match,' the number of non-stopword tokens in the response that are present in the document but not present in the context of the conversation. Excluding words from the conversation history means that, in order to produce a word of the document, the response generation system is very likely to be effectively influenced by that document. We then compute both *precision* as '#match' divided by the total number of non-stop tokens in the response, and *recall* as '#match' divided by the total number of non-stop tokens in the document. We also compute the respective *F1* score to combine both. Looking only at exact unigram matches between the document and response is a major simplifying assumption, but the combination of the three metrics offers a plausible proxy for how greatly the response is grounded in the document. It seems further reasonable to assume that these can serve as a surrogate for less quantifiable forms of grounding such as paraphrase – e.g., *US* → *American* – when the statistics are aggregated on a large test dataset.

6.2 Automatic Evaluation

Table 2 shows automatic evaluation results for the different systems. In terms of appropriateness, the different variants of our models outperform the SEQ2SEQ and MEMNET baselines, but differences are relatively small and, in case of one of the metrics (NIST), the best system does not use grounding. Our goal, we would note, is not to specifically improve response appropriateness, as many responses that completely ignore the document (e.g., *I don't know*) might be perfectly appropriate. Our systems fare much better in terms of Grounding and Diversity: our best system (CMR+W) achieves an F1 score that is more than three times (0.38% vs. 0.12%) higher than the most competitive non-MRC system (MEMNET).

Human judges preferred:				
Our best system		Neutral	Comparator	
CMR+W	*44.17%	26.27%	29.56%	SEQ2SEQ
CMR+W	*40.93%	25.80%	33.27%	MEMNET
CMR+W	37.67%	27.53%	34.80%	CMR
CMR+W	30.37%	16.27%	*53.37%	Human

Table 3: **Human Evaluation** results, showing preferences (%) for our model (CMR+W) vs. baseline and other comparison systems. Distributions are skewed towards CMR+W. The 5-point Likert scale has been collapsed to a 3-point scale. *Differences in mean preferences are statistically significant ($p \leq 0.0001$).

6.3 Human Evaluation

We sampled 1000 conversations from the test set. Filters were applied to remove conversations containing ethnic slurs or other offensive content that might confound judgments. Outputs from systems to be compared were presented pairwise to judges from a crowdsourcing service. Four judges were asked to compare each pair of outputs on Relevance (the extent to which the content was related to and appropriate to the conversation) and Informativeness (the extent to which the output was interesting and informative). Judges were asked to agree or disagree with a statement that one of the pair was better than the other on the above two parameters, using a 5-point Likert scale.[9] Pairs

[9]The choices presented to the judges were *Strongly Agree*, *Agree*, *Neutral*, *Disagree*, and *Strongly Disagree*.

of system outputs were randomly presented to the judges in random order in the context of short snippets of the background text. These results are presented in summary form in Table 3, which shows the overall preferences for the two systems expressed as a percentage of all judgments made. Overall inter-rater agreement measured by Fliess' Kappa was 0.32 ("fair"). Nevertheless, the differences between the paired model outputs are statistically significant (computed using 10,000 bootstrap replications).

6.4 Qualitative Study

Table 4 illustrates how our best model (CMR+W) tends to produce more contentful and informative responses compared to the other systems. In the first example, our system refers to a particular *episode* mentioned in the article, and also uses terminology that is more consistent with the article (e.g., *series*). In the second example, *humorous song* seems to positively influence the response, which is helpful as the input doesn't mention singing at all. In the third example, the CMR+W model clearly grounds its response to the article as it states the fact (Steve Jobs: CEO of Apple) retrieved from the article. The outputs by the other two baseline models are instead not relevant in the context.

Figure 3 displays the attention map of the generated response and (part of) the document from our full model. The model successfully attends to the key words (e.g., *36th, episode*) of the document. Note that the attention map is unlike what is typical in machine translation, where target words tend to attend to different portions of the input text. In our task, where alignments are much less one-to-one compared to machine translation, it is common for the generator to retain focus on the key information in the external document to produce semantically relevant responses.

7 Related Work

Dialogue: Traditional dialogue systems (see (Jurafsky and Martin, 2009) for an historical perspective) are typically grounded, enabling these systems to be reflective of the user's environment. The lack of grounding has been a stumbling block for the earliest end-to-end dialogue systems, as various researchers have noted that their outputs tend to be bland (Li et al., 2016a; Gao et al., 2019b), inconsistent (Zhang et al., 2018a; Li et al., 2016b; Zhang et al., 2019), and lacking in factual content (Ghazvininejad et al., 2018; Agarwal et al., 2018). Recently there has been growing interest in exploring different forms of grounding, including images, knowledge bases, and plain texts (Das et al., 2017; Mostafazadeh et al., 2017; Agarwal et al., 2018; Yang et al., 2019). A recent survey is included in Gao et al. (2019a).

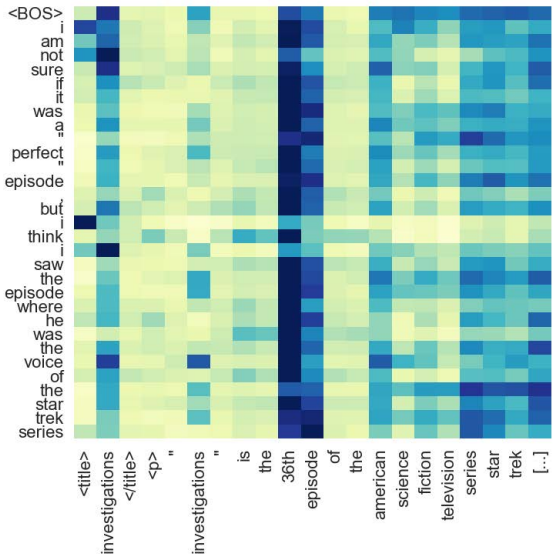

Figure 3: Attention weights between words of the documents and words of the response. Dark (blue) cells represent probabilities closer to 1.

Prior work, e.g, (Ghazvininejad et al., 2018; Zhang et al., 2018a; Huang et al., 2019), uses grounding in the form of independent snippets of text: Foursquare tips and background information about a given speaker. Our notion of grounding is different, as our inputs are much richer, encompassing the full text of a web page and its underlying structure. Our setting also differs significantly from relatively recent work (Dinan et al., 2019; Moghe et al., 2018) exploiting crowdsourced conversations with detailed grounding labels: we use Reddit because of its very large scale and better characterization of real-world conversations. We also require the system to learn grounding directly from conversation and document pairs, instead of relying on additional grounding labels. Moghe et al. (2018) explored directly using a span-prediction QA model for conversation. Our framework differs in that we combine MRC models with a sequence generator to produce free-form responses.

Machine Reading Comprehension: MRC models such as SQuAD-like models, aim to extract answer spans (starting and ending indices)

Doc	<title> Investigations </title> "Investigations" is the **36th episode** of the American science fiction **television series** Star Trek: Voyager which aired on the UPN network. It is the **36th episode** of the second season. Set in the 24th century, **the series** follows the adventures of the Federation starship Voyager during its journey home to Earth, having been stranded tens of thousands of light-years away. [...]
Input	[...] King Abdullah II of Jordon once cameo'd as an unnamed ensign on Star Trek Voyager
MEMNET	I think it should be noted that he is like a good guy who just killed her and her son.
SEQ2SEQ	well , i didn't know , but i was just saying that it wasn't a joke.
CMR+W	I am not sure if it was a **"perfect" episode**, but I think I saw the **episode** where he was the voice of the **Star Trek series**.
Doc	<title> "Weird Al" Yankovic </title> Alfred Matthew "Weird Al" Yankovic is an American singer, songwriter, parodist, record producer, satirist, actor, voice actor, music video director, film producer, and author. He is **known** for **his humorous songs** that make light of popular culture and often parody specific songs by contemporary musical acts, original songs that are style pastiches of the work of other acts, and polka medleys of several popular songs, featuring his favored instrument, the accordion. *(***Prince** *and* **West** *mentioned several times in the rest of the article.)* [...]
Input	That's fair, I believe Weird Al always got permission from artists before he parodied them which is nice.
MEMNET	Well then you don't need to make a sequel to the front page of the movie.
SEQ2SEQ	If I remember correctly. I had no idea who had any good chance in history class.
CMR+W	No, because **his song was a parody** of the artist **formerly known as "Kanye West"**.
Doc	<title> Steve Jobs </title> Steven Paul Jobs was an American entrepreneur, businessman, inventor, and industrial designer. He was the chairman, **chief executive officer (CEO)**, and co-founder of **Apple Inc**.; chairman and majority shareholder of Pixar; a member of The Walt Disney Company's board of directors following its acquisition of Pixar; and the founder, chairman, and CEO of next. [...]
Input	Steve Jobs was a mediocre programmer and one of the greatest designers and marketers in human history. But I prefer Bill Gates if we want to talk about hero worship myself. [...]
MEMNET	What if you don't want to see this movie?
SEQ2SEQ	No ... the whole point is that this is a good way to make a difference.
CMR+W	So he's the **CEO** of **Apple**.

Table 4: Sample output comparing our best system (CMR+W) against Memory Networks and a SEQ2SEQ baseline. The source documents were manually shortened to fit in the table, without significantly affecting meaning.

from a given document for a given question (Seo et al., 2017; Liu et al., 2018b; Yu et al., 2018). These models differ in how they fuse information between questions and documents. We chose SAN (Liu et al., 2018b) because of its representative architecture and competitive performance on existing MRC tasks. We note that other off-the-shelf MRC models, such as BERT (Devlin et al., 2018), can also be plugged in. We leave the study of different MRC architectures for future work. Questions are treated as entirely independent in these "single-turn" MRC models, so recent work (e.g., CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018)) focuses on multi-turn MRC, modeling sequences of questions and answers in a conversation. While multi-turn MRC aims to answer complex questions, that body of work is restricted to factual questions, whereas our work—like much of the prior work in end-to-end dialogue—models free-form dialogue, which also encompasses chitchat and non-factual responses.

8 Conclusions

We have demonstrated that the machine reading comprehension approach offers a promising step to generating, *on the fly*, contentful conversation exchanges that are grounded in extended text corpora. The functional combination of MRC and neural attention mechanisms offers visible gains over several strong baselines. We have also formally introduced a large dataset that opens up interesting challenges for future research.

The CMR (Conversation with on-demand machine reading) model presented here will help connect the many dots across multiple data sources. One obvious future line of investigation will be to explore the effect of other off-the-shelf machine reading models such as BERT (Devlin et al., 2018) within the CMR framework.

Acknowledgements

We are grateful to the anonymous reviewers, as well as to Vighnesh Shiv, Yizhe Zhang, Chris Quirk, Shrimai Prabhumoye, and Ziyu Yao for helpful comments and suggestions on this work. This research was supported in part by NSF (IIS-1524371), DARPA CwC through ARO (W911NF-15-1-0543), and Samsung AI Research.

References

Shubham Agarwal, Ondrej Dusek, Ioannis Konstas, and Verena Rieser. 2018. A knowledge-grounded multimodal search-based conversational agent. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 59–66, Brussels, Belgium.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. of HLT*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proc. of EMNLP*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proc. of ACL*.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded response generation task at DSTC7. In *AAAI Dialog System Technology Challenges Workshop*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2019a. Neural approaches to conversational ai. *Foundations and Trends in Information Retrieval*, 13(2-3):127–298.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019b. Jointly optimizing diversity and relevance in neural response generation. In *NAACL-HLT 2019*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proc. of AAAI*.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2019. Challenges in building intelligent open-domain dialog systems. *arXiv preprint arXiv:1905.05709*.

Dan Jurafsky and James H Martin. 2009. *Speech & language processing*. Prentice Hall.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proc. of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proc. of NAACL-HLT*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proc. of ACL*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proc. of EMNLP*.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018a. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1489–1498.

Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2018b. Stochastic Answer Networks for machine reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1694–1704, Melbourne, Australia.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6297–6308.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proc. of EMNLP*.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proc. of IJCNLP*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association of Computational Linguistics (TACL)*.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *ICLR*.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. *AAAI*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proc. of ACL-IJCNLP*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proc. of NAACL-HLT*.

Sainbayar Sukhbaatar, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Proc. of NIPS*.

Ilya Sutskever, Oriol Vinyals, and Quoc Le. 2014. Sequence to sequence learning with neural networks. In *Proc. of NIPS*, pages 3104–3112.

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 *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *ICML Deep Learning Workshop*.

Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A hybrid retrieval-generation neural conversation model. *arXiv preprint arXiv:1904.09068*.

Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, R. Chulaka Gunasekara, Walter S. Lasecki, Jonathan K. Kummerfeld, Michel Galley, Chris Brockett, Jianfeng Gao, Bill Dolan, Xiang Gao, Huda AlAmri, Tim K. Marks, Devi Parikh, and Dhruv Batra. 2019. Dialog system technology challenge 7. In *In NeurIPS Conversational AI Workshop*.

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 *ICLR*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. In *Proc. of NeurIPS*.

Yizhe Zhang, Xiang Gao, Sungjin Lee, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Consistent dialogue generation with self-supervised feature learning. *arXiv preprint arXiv:1903.05759*.