Towards Exploiting Background Knowledge for Building Conversation Systems

Nikita Moghe[1,2], **Siddhartha Arora**[1], **Suman Banerjee**[1], **and Mitesh M. Khapra**[1,2]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Madras
[2]Robert Bosch Centre for Data Science and AI (RBC-DSAI),
Indian Institute of Technology Madras
{nikitavam,sidarora,suman,miteshk}@cse.iitm.ac.in

Abstract

Existing dialog datasets contain a sequence of utterances and responses without any explicit background knowledge associated with them. This has resulted in the development of models which treat conversation as a sequence-to-sequence generation task (*i.e.*, given a sequence of utterances generate the response sequence). This is not only an overly simplistic view of conversation but it is also emphatically different from the way humans converse by heavily relying on their background knowledge about the topic (as opposed to simply relying on the previous sequence of utterances). For example, it is common for humans to (involuntarily) produce utterances which are copied or suitably modified from background articles they have read about the topic. To facilitate the development of such natural conversation models which mimic the human process of conversing, we create a new dataset containing movie chats wherein each response is explicitly generated by copying and/or modifying sentences from unstructured background knowledge such as plots, comments and reviews about the movie. We establish baseline results on this dataset (90K utterances from 9K conversations) using three different models: (i) pure generation based models which ignore the background knowledge (ii) generation based models which learn to copy information from the background knowledge when required and (iii) span prediction based models which predict the appropriate response span in the background knowledge.

1 Introduction

Background knowledge plays a very important role in human conversations. For example, to have a meaningful conversation about a movie, one uses their knowledge about the plot, reviews, comments and facts about the movie. A typical conversation involves recalling important points from this background knowledge and producing them appropriately in the context of the conversation. However, most existing large scale datasets (Lowe et al., 2015b; Ritter et al., 2010; Serban et al., 2016) simply contain a sequence of utterances and responses without any *explicit* background knowledge associated with them. This has led to the development of models which treat conversation as a simple sequence-to-sequence generation task and often produce output which is both syntactically incorrect and incoherent (off topic). To make conversations more coherent, there is an increasing interest in integrating structured and unstructured knowledge sources with neural conversation models. While there are already some works in this direction (Rojas-Barahona et al., 2017; Williams et al., 2016; Lowe et al., 2015a; Ghazvininejad et al., 2017) which try to integrate external knowledge sources with existing datasets, we believe that building new datasets where the utterances are *explicitly* linked to external background knowledge will further facilitate the development of such background aware conversation models.

With this motivation, we built a new background aware conversation dataset using crowdsourcing. Specifically, we asked workers to chat about a movie using structured and unstructured resources about the movie such as plots, reviews, comments, fact tables (see Figure 1). For every even numbered utterance, we asked the workers to consult the available background knowledge and try to construct a sentence which contains information from this background knowledge and is relevant in the current context of the conversation (akin to how humans recall things from their background knowledge and insert them appropriately in the conversation). For example, in Turn 2, Speaker 2 picked a sentence from the plot which is relevant to the current context of the conversation. Similarly, in Turn 3, Speaker 2 picked a

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2322–2332
Brussels, Belgium, October 31 - November 4, 2018.

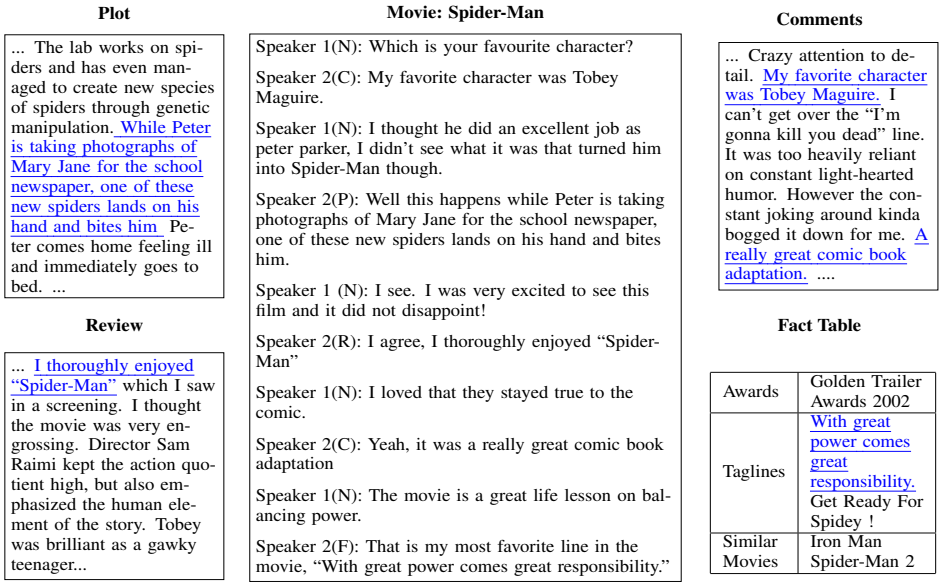

Figure 1: A sample chat from our dataset which uses background resources. The chosen spans used in the conversation are shown in blue. The letters in the brackets denote the type of resource that was chosen - P, C ,R, F and N indicate **P**lot, **C**omments, **R**eview, **F**act Table and **N**one respectively.

sentence from the movie review. We also asked the workers to suitably modify the content picked from the background knowledge, if needed, so that the conversation remains coherent. We collected around 9K such conversations containing a total of 90K utterances pertaining to about 921 movies. These conversations along with the background resources will be made publicly available[1]. For every utterance, we also provide information about the exact span in the resource from which this utterance was created. Lastly note that unlike existing datasets, our *test set* contains multiple reference responses for each test context thereby facilitating better evaluation of conversation models. We believe that this dataset will allow the community to take a fresh look at conversation modeling and will lead to the development of models which can learn to exploit background knowledge to pick appropriate responses instead of generating responses from scratch. Such a conversation strategy which produces responses from background knowledge would be useful in various domains. For example, a troubleshooting bot could exploit the information available in manuals, reviews and previous bug reports about the software. Similarly, an e-commerce bot could exploit the rich information available in product descriptions, reviews, fact tables, *etc.* about the product. While the proposed dataset is domain specific, it serves as a good benchmark for developing creative background-knowledge-aware models which can then be ported to different domains by building similar datasets for other domains.

We establish some initial baselines using three different paradigms to demonstrate the various models that can be developed and evaluated using this dataset. For the sake of completeness, the first paradigm is a hierarchical variant of the sequence to sequence architecture which does not exploit any background knowledge. The second paradigm is the copy-and-generate paradigm wherein the model tries to copy text from the given resources whenever appropriate and generate it otherwise. The third paradigm borrows from the span prediction based models which are predominantly being used for Question Answering (QA). These baseline results along with the dataset would hopefully shape future research in the area of background aware conversation models.

2 Related Work

There has been an active interest in building datasets (Serban et al., 2015) for training dialog systems. Some of these datasets contain transcripts of human-bot conversations (Williams et al., 2013; Henderson et al., 2014a,b) while others are created using a fixed set of natural language patterns (Bordes and Weston, 2017; Dodge et al., 2016). The advent of deep learning created

[1] https://github.com/nikitacs16/Holl-E

interest in the construction of large-scale dialog datasets (Lowe et al., 2015b; Ritter et al., 2010; Sordoni et al., 2015) leading to the development of several end-to-end conversation systems (Shang et al., 2015; Vinyals and Le, 2015; Li et al., 2016; Serban et al., 2016) which treat dialog as a sequence generation task.

To make the output of these models more coherent, there is an increasing effort in integrating external background knowledge with these models. This is because human beings rely on background knowledge for conversations as well as other tasks (Schallert, 2002). There has been considerable work on incorporating background knowledge in the context of goal-oriented dialog datasets even before the advent of large-scale datasets for deep learning (Raux et al., 2005; Seneff et al., 1991) as well as in recent times (Rojas-Barahona et al., 2017; Williams et al., 2016; Eric et al., 2017) where datasets include small sized knowledge graphs as background knowledge. However, the conversations in these datasets are very templated and nowhere close to open conversations in specific domains such as the ones contained in our dataset.

Even in the case of open domain conversations, there are some works which have integrated external knowledge sources. Most of the entries in 2017 Amazon Alexa Prize (Ram et al., 2017) relied on background knowledge for meaningful response generation. Milabot (Serban et al., 2017a) and even the winning entry SoundingBoard (Liu et al., 2018) used Reddit pages, Amazon's Evi Service, and large databases like OMDB, Google Knowledge Graph and Wikidata as external knowledge. The submission named Eigen (Guss et al., 2017) used several dialog datasets and corpora belonging to related Natural Language Processing tasks to make their responses more informative. We refer the reader to (Ram et al., 2017) for detailed analysis of these systems. In the space of academic datasets, Lowe et al. (2015a) report results on the Ubuntu dataset using manpages as external knowledge whereas Ghazvininejad et al. (2017) use Foursquare tips as external knowledge for social media conversations. However, unlike our work both these works do not create a new dataset where the responses are explicitly linked to a knowledge source. The infusion of external knowledge in both these works is post facto (as opposed to our work where we take a bottom-up approach and explicitly create a dataset which allows exploitation of background knowledge). Additionally, existing large-scale datasets are noisy as they are extracted from online forums which are inherently noisy. In contrast, since we use crowdsourcing, the extent of noise is reduced since there are humans in the loop who were explicitly instructed to use only clean sentences from the external knowledge sources.

We would also like to mention some existing works such as (He et al., 2017; Lewis et al., 2017; Krause et al., 2017) which have used crowdsourcing for creating conversation datasets. In fact, our data collection method is inspired by the work of Krause et al. (2017) where the authors use self-dialogs to collect conversation data about movies, music and sports. They are referred to as self-dialogs because the same worker plays the role of both parties in the conversation. However, our work differs from Krause et al. (2017) as we provide explicit background knowledge sources to the workers from where they can copy text with the addition of suitable prefixes and suffixes to generate appropriate coherent responses.

3 Dataset

In the following sub-sections we describe the various stages involved in collecting our dataset.

3.1 Curating a list of popular movies

We created a list of 921 movies containing (i) top 10 popular movies within the past five years, (ii) top 250 movies as per IMDb rankings, (iii) top 10 movies in popular genres, and (iv) other popular movie lists made available elsewhere on the Internet. These movies belonged to 22 different genres such as sci-fi, action, horror, fantasy, adventure, romance, *etc.* thereby ensuring that our dataset is not limited to a specific genre. We considered those movies for which enough background information such as plots, reviews, comments, facts, *etc.* were available on the Internet irrespective of whether they were box-office successes or not. Please find the respective urls in the Appendix.

3.2 Collecting background knowledge

For each movie, we collected the following background knowledge:

1. Review (R): For each movie, we asked some in-house workers to fetch the top 2 most popular reviews for this movie from IMDb using the *sort*

by Total Votes option. We also instructed them to avoid choosing reviews which were less than 50 words but this was typically never the case with popular reviews. **2. Plot (P):** For each movie, we extracted information about the "Plot" of the movie from the Wikipedia page of the movie. Wikipedia pages of movies have an explicit section on "Plot" making it easy to extract this information using scripts. **3. Comments (C):** Websites like *Reddit* have a segment called "official discussion page about X" (where X is a movie name) containing small comments about various aspects of movie. We identified such pages and extracted the first comment on every thread on this page. We bundled all these comments into a single text file and refer to it as the resource containing "Comments". For a few movies, the official discussion page was not present in which case we used the review titles of all the IMDb reviews of the movie as comments. The difference between Reviews and Comments is that a Review is an opinion piece given by one person thus typically exhibiting one sentiment throughout while Comments include opinions of several people about the same movie ensuring that positive, negative and factual aspects of the movie are captured as well as some banter.

4. Meta data or Fact Table (F): For each movie, we also collected factual details about the movie, *viz.*, box office collection, similar movies (for recommendations), awards and tag-lines from the corresponding IMDb pages and Wikipedia Infoboxes. Such information would be useful for inserting facts in the conversation, for example, *"Did you know that the movie won an Oscar?*". We included only 4 fields in our fact table instead of showing the entire Wikipedia Infobox to reduce the cognitive load on turkers who already had to read the plot, reviews and comments of the movie.

3.3 Collecting conversation starters

During our initial pilots, we observed that if we asked the workers to converse for at least 8 turns, they used a lot of the initial turns in greetings and general chit-chat before actually chatting about a movie. To avoid this, we collected opening statements using Amazon Mechanical Turk (AMT) where the task for the workers was to answer the following questions "*What is your favorite scene from the movie X ?*", "*What is your favorite character from the movie X ?*" and "*What is your opinion about the movie X?*" (X is the movie name). We paid the workers 0.04$ per movie and showed the same movie to 3 different workers, thereby collecting 9 different opening statements for every movie. By using these statements as conversation starters in our data collection, the workers could now directly start conversing about the movie.

3.4 Collecting background knowledge aware conversations via crowdsourcing

Our aim is to create a conversation dataset wherein every response is explicitly linked to some structured or unstructured background knowledge. Creating such a dataset using dedicated in-house workers would obviously be expensive and time consuming and so we decided to use crowdsourcing. However, unlike other NLP and Vision tasks, where crowdsourcing has been very successful, collecting conversations via crowdsourcing is a bit challenging. The main difficulty arises from the fact that conversation is inherently a task involving two persons but it is hard to get two workers to synchronize and chat on AMT. We did try a few pilot experiments where we setup a server to connect two AMT workers but we found that the probability of two workers simultaneously logging in was very low. Thus, most workers logged in and left in a few seconds because no other worker joined simultaneously. Finally, we took inspiration from the idea of self chats Krause et al. (2017) in which, the same worker plays the role of both Speaker 1 and Speaker 2 to create the chat. In the above self chat setup, we showed every worker 3 to 4 resources related to the movie, *viz.*, plot (P), review (R), comments (C) and fact table (F). We also showed them a randomly selected opening statement from the 9 opening statements that we had collected for each movie and requested them to continue the conversation from that point. The workers were asked to add at least 8 utterances to this initial chat. While playing the role of Speaker 1, the worker was not restricted to copy/modify sentences from the background resources but was given the freedom to create (write) original sentences. However, when playing the role of Speaker 2, the worker was strictly instructed to copy/modify sentences from the shown resources such that they were relevant in the current context of the conversation. The reason for not imposing any restrictions on Speaker 1 was to ensure that the chats look more natural and

coherent. Further, Speaker 2 was allowed to add words at the beginning or end of the span selected from the resources to make the chats more coherent and natural (for example, see the prefix in utterance 2 of Speaker 2 in Figure 1). We paid the workers 40 cents for every chat. Please refer to the Appendix for the instruction screen shots.

3.5 Verification of the collected chats

Every chat that was collected by the above process was verified by an in-house evaluator to check if the workers adhered to the instructions and produced coherent chats. Since humans typically tend to paraphrase the background knowledge acquired by reading articles, one could argue that such conversations may not look very natural because of this restriction to copy/modify content from the provided resources. To verify this, we conducted a separate human evaluation wherein we asked 15 in-house evaluators to read conversations (without the background resources) from our dataset and rate them on five different parameters. Specifically, they were asked to check if the conversations were 1) **intelligible:** *i.e.*, an average reader could understand the conversation 2) **coherent:** *i.e.*, there were no abrupt context switches 3) **grammatically correct** 4) **on-topic:** *i.e.*, the chat revolved around the concerned movie with digression limited to related movies/characters/actors and 5) **natural two-person chats:** *i.e.*, the role-play setup does not make the chat look unnatural. These evaluators were post-graduate students who were fluent in English and had watched at least 100 Hollywood movies. We did not give them any information about the data creation process. We used a total of 500 chats for the evaluation and every chat was shown to 3 different evaluators. The evaluators rated the conversations on a scale of 1 (very poor) to 5 (very good). We computed inter-annotator agreement using the mean linearly weighted Cohen's κ (Cohen, 1968) and mean Krippendorff's α (Hayes and Krippendorff, 2007). The average rating for each of the 5 parameters along with the inter annotator agreement are reported in Table 1 and are very encouraging.

3.6 Statistics

In Table 2, we show different statistics about the dataset collected using the above process. These include average number of utterances per chat, average number of words per utterance, and so on followed by the statistics of the different resources which were used as background knowledge. Please note that the # unique Plots and # unique Reviews correspond to unique paragraphs while the # unique Comments is the count of unique sentences. We observed that 41.2%, 34.6%, 16.1% and 8.1% of Speaker 2 responses came from Reviews, Comments, Plots and Fact Table respectively.

Metric	Rating	α	κ
Intelligible	4.47 ± 0.52	0.70	0.69
Coherent	4.33 ± 0.93	0.57	0.71
Grammar	4.41 ± 0.56	0.60	0.69
Two-person-chat	4.47 ± 0.46	0.64	0.70
On Topic	4.57 ± 0.43	0.72	0.70

Table 1: Average human evaluation scores with standard deviations for conversations (scale 1-5). We also report mean Krippendorff's α and mean Cohen's κ

#chats	9071
#movies	921
#utterances	90810
Average # of utterances per chat	10.01
Average # of words per utterance	15.29
Average # of words per chat	153.07
Average # of words in Plot	186.10
Average # of words in Review	384.44
Average # of words in Comments	123.81
Average # of words in Fact Table	33.47
# unique Plots	5157
# unique Reviews	1817
# unique Comments	12740

Table 2: Statistics of the dataset

4 Models

We evaluate three different types of models as described below. Since these are popular existing models, we describe them very briefly below and refer the reader to the original papers for more details. Note that in this work we merge the comments, reviews, plots and facts into one single document and refer to it as background knowledge. In the rest of the paper, when we refer to a *resource* we mean this single document which is a merger of all the resources unless specified otherwise.

4.1 Generation based models

We use the standard Hierarchical Recurrent Encoder Decoder model (HRED) (Serban et al., 2016) instead of its variant (Serban et al., 2017b)

as the standard model performs only slightly poorly than the variant and is much easier to implement. It decomposes the context of the conversation as two level hierarchy using Recurrent Neural Networks (RNN). The lower RNN encodes individual utterances (sequence of words) which is then fed into the higher level RNN as a sequence of utterances. The decoder RNN then generates the output based on this hierarchical context representation.

4.2 Generate-or-Copy models

Get To The Point (GTTP) (See et al., 2017) proposed a hybrid pointer generator network for abstractive summarization that learns to copy words from the source document when required and otherwise generates a word like any sequence-to-sequence model. In the summarization task, the input is a *document* and the output is a *summary* whereas in our case the input is a {*document, context*} pair and the output is a *response*. Here, the context includes the previous two utterances and the current utterance. We modified the architecture to suit our task. We use an RNN to compute the representation of the document (like the original model) and introduce another RNN to compute a representation of the context by treating it as a single sequence of words. The decoder which is also an RNN then uses the document representation, context representation and its own internal state representation to compute a (i) probability score which indicates whether the next word should be copied or generated (ii) probability distribution over the vocabulary if the next word needs to be generated and (iii) probability distribution over the input words if the next word needs to be copied. These three probability distributions are then combined to produce the next word in the response.

4.3 Span prediction models

Bi-directional Attention Flow Model (BiDAF) (Seo et al., 2017) model is a QA model which was proposed in the context of the SQuAD dataset (Rajpurkar et al., 2016). Given a *document* and a *question*, the model uses a six-layered architecture to predict the span in the document which contains the answer. We can use their model as it is for our task without any modifications by simply treating the *context* as the *question* and the *resource* as the *document*.

We chose to evaluate on the modified generate-or-copy model instead of other variants such as (Ghazvininejad et al., 2017; Lowe et al., 2015a) as the modified model already contains the extra encoder for background model which is present in these models. Moreover, the modified model uses a hybrid copy-or-generate decoder which is well-suited to our task.

5 Experimental Setup

In this section we describe the train-validation-test splits, the process used for creating training instances, the manner in which the models were trained using our data and the evaluation metrics.

5.1 Creating train/valid/test splits

On average we have 9.14 chats per movie. We divide the collected chats into train, validation, and test splits such that all the chats corresponding to a given movie are in exactly one of the splits. This ensures that a movie seen in the test or validation set is never seen at training time. We create the splits such that the percentage of chats in the train-validation-test set is roughly 80%-10%-10%.

5.2 Creating training instances

For each chat in the training data, we construct training instances of the form {*resource, context, response*} where the *context* is taken as previous two utterances and current utterance. We consider only the even numbered utterances as training examples as they are generated from the background resources thus emulating a human-bot setup. If a chat has 10 turns, we will have 5 instances. The task then is to train a model which can predict these even numbered responses. At test time the model is shown {*resource, context*} and predicts the response. Note that, HRED will ignore the *resource* and only use {*context, response*} as input-output pairs. BiDAF and GTTP will use {*resource, context, response*} as training data with relevant *span* instead of *response* for BiDAF.

5.3 Merging resources into a single document

As stated earlier, we simply merge all the background information to create a single document which we collectively refer to as *resource*. For the BiDAF model, we had to restrict the length of the resource to 256 words because we found that even on a K80 GPU with 12GB RAM, this model gives an out of memory error for longer

documents. We found this to be a severe limitation of this and other span based models (for example, R-Net (Wang et al., 2017)) . We experimented with three methods of creating this resource. The first method *oracle* uses the actual resource (plot or comments or reviews) from which the next response was generated as a resource. If that resource itself has more than 256 words then we truncate it from the beginning and the end such that the span containing the actual response is contained within the retained 256 words. The number of words that are discarded from the start or the end is chosen at random so that the correct spans do not end up in similar positions throughout the dataset. The next two methods *mixed-short* and *mixed-long* are created by merging the individual resources. We retain each resource in the merged document proportional to its length. (*i.e,*if there are 400 words in the plot, 200 words in the review and 100 in the comments, the merged resource will contain contiguous sentences from these three resources in the ratio of 4:2:1.) Further, we ensure that the merged resource contains the actual response span. In this way, we create *mixed-short* with 256 words and *mixed-long* with 1200 words (the maximum length of the merged resources). We will henceforth denote *oracle*, *mixed-long* and *mixed-short* using '(o) ', '(ms) 'and '(ml) 'respectively. We report results for BiDAF(o), BiDAF (ms), GTTP (o) and GTTP (ml).

5.4 Evaluation metrics

As HRED and GTTP models are generation based models we use BLEU-4, ROUGE-1, ROUGE-2 and ROUGE-L as the evaluation metrics. For BiDAF we use the above metrics by comparing the predicted span with the reference span. For BiDAF, we also report F1 as stated in Rajpurkar et al. (2016).

In addition to the automatic evaluation, we also collected human judgments using 100 test responses generated for every model for every setup (oracle, mixed-short, mixed-long). These evaluators had the same qualifications as the evaluators who earlier helped us evaluate our dataset. They were asked to rate the response on scale of 1 to 5 (with 1 being the least) on the following four metrics: (1) Fluency(Flu), (2) appropriateness/relevance (apt) of the response in the current context language (3) humanness (Hum) of the response, *i.e.,* whether the responses look as if they were generated by a human (4) and specificity (spec) of the response, *i.e.*, whether the model produced movie-specific responses or generic responses such as "This movie is amazing". We report these results in Table 4.

5.5 Collecting multiple reference responses

One common issue with evaluating dialog systems is that existing datasets typically contain only one reference response whereas in practice several responses can be correct in a given context. To solve this to a certain extent, we collected three reference responses for every Speaker 2 utterance in our dataset (note that Speaker 2 is treated as the bot while training/testing our models). We show the previous utterances ending with Speaker 1's response and ask workers to provide three appropriate responses from the given resources. We found that in some cases there was only one appropriate response like factual response and the workers could not provide multiple references . In this way we were able to create a multiple reference test set where 78.04% of the test instances have multiple responses. In Table 3, we report two sets of scores based on single-reference test dataset and multi-reference test dataset. While calculating the scores for multi-reference dataset, we take the maximum score over multiple reference responses.

Please refer to the Appendix section for the details of the model, hyperparameters, example of multiple references in our dataset and sample outputs produced by different models.

6 Results and Discussion

In this section, we discuss the results of our experiments as summarized in Tables 3 and 4.

Generation based models v/s Span prediction models: We compare the generation based models and span prediction models only based on results in the *oracle* setting. Here, the span based model (BiDAF) outperforms the generation based models (HRED and GTTP). This confirms our belief that the natural language generation (NLG) capabilities of current generation based models are far from being acceptable even in case of generate-or-copy modes. This also emphasizes the importance of this data which allows building models which can exploit well-formed sentences in the background knowledge and reproduce them with minor modifications instead of generating them from scratch. While the results for BiDAF are

Model	F1		BLEU		Rouge-1		Rouge-2		Rouge-L	
HRED	-	-	5.23	5.38	24.55	25.38	7.61	8.35	18.87	19.67
GTTP (o)	-	-	13.92	16.46	30.32	31.6	17.78	21.21	25.67	27.83
GTTP (ms)	-	-	11.05	15.68	29.66	31.71	17.70	19.72	25.13	27.35
GTTP (ml)	-	-	7.51	8.73	23.20	21.55	9.91	10.42	17.35	18.12
BiDAF (o)	39.69	47.18	28.85	34.98	39.68	46.49	33.72	40.58	35.91	42.64
BiDAF (ms)	45.73	51.35	32.95	39.39	45.69	50.73	40.18	45.01	43.46	46.95

Table 3: Performance of the proposed models on our dataset. The figures on the left in each column indicate scores on single-reference test dataset while the figures on the right denote scores on multi-reference dataset.

Model	Hum	Apt	Flu	Spec
HRED	3.08	2.49	2.64	2.06
GTTP (o)	4.10	3.73	4.03	3.33
GTTP (ml)	2.93	2.97	3.42	2.60
BiDAF (o)	3.78	3.71	4.05	3.76
BiDAF(ms)	3.41	3.38	3.47	3.30

Table 4: Human evaluation results on the model performances.

encouraging, we reiterate that it does not scale to longer documents (we were not able to run it in the *mixed-long* setting). We still need much better models as BiDAF on SQuAD dataset gives an F1 of 81.52 % which is much higher than the results on our dataset. Further, note that using the predicted span as a response is not natural. This is evident from human likeliness (Hum) score of GTTP (o) being higher than both the BiDAF models. We need models which can suitably alter the span to retain the coherence of the context.

Effect of including background knowledge: We observe that there isn't much difference between the performance of HRED which does not use any background knowledge when compared to GTTP (ml) which actually uses a lot of background knowledge. However, there is a substantial difference between the performance of HRED and GTTP (o) which uses only the relevant background knowledge. Further, without background knowledge, HRED learns to produce very generic responses (Spec score = 2.06). This shows that the background knowledge is important, but the models should learn to focus on the right background knowledge relevant to the current context. Alternately, we can have a two-stage network which first predicts the right resource (plot, review, comments) from which the span should be selected and then selects the span from this chosen resource.

***Oracle* v/s *mixed-short* resource:** We observe that the performance of BiDAF (ms) is actually better than BiDAF (o) even when the resource length for both is 256 words. We would expect a poor performance for BiDAF (ms) as the resource has more noise because of the sentences from irrelevant resources. However, we speculate the model learns to regard irrelevant sentences as noise and learns to focus on sentences corresponding to the correct resource resulting in improved performance (however, this is only a hypothesis and it needs to be verified). We realize that this is clearly a poor baseline and we need better span prediction based models which can work with longer documents. At the same time, GTTP (o) and GTTP (ms) have comparable (yet poor) performance. There is no co-attention mechanism in this model which can effectively filter out noisy sentences.

Observations from the copy-and-gen model: We observed that this model produced sentences where on average of 82.18% (*oracle*) and 71.95% (*mixed-long*) of the tokens were copied. One interesting observation was that it easily learns to copy longer contiguous sequences one word at a time. However, as is evident from the automatic evaluation metrics, in many cases, the 'copied' spans are not relevant to the current context.

Evaluating with multiple references: When considering multiple references, the performance numbers as reported in Table 3 indeed improve. This shows the importance of having multiple references and the need to develop metrics which account for multiple dissimilar references.

7 Conclusion

We introduce a new dataset for building dialog systems which would hopefully allow the community to take a fresh look at this task. Unlike existing datasets which only contain a sequence of utterances, in our dataset each response is explic-

itly linked to some background knowledge. This mimics how humans converse by recalling information from their background knowledge and use it appropriately in the context of the conversation. Using this dataset, we evaluated models belonging to three different paradigms, *viz.*, generation based models, generate-or-copy models and span prediction models. Our results suggest that the NLG capabilities of existing seq-to-seq models are still far from desirable while span based models which completely bypass the process of NLG show some promise but with clear scope for improvement.

Going forward, we would like to build models which are a hybrid of span prediction models and generation models. Specifically, we would like to build models which can learn to copy a large sequence from the input instead of one word at a time. Another important aspect is to build less complex models which can handle longer documents. For example, the BiDAF model has an expensive outer product between two large matrices which makes it infeasible for long documents (because the size of these matrices grows with the length of the document). Alternately, we would like to build two-stage models which first select the correct resource from which the next response is to be generated and then generate or copy the response from the resource.

Acknowledgements

We would like to thank Department of Computer Science and Engineering, and Robert Bosch Center for Data Sciences and Artificial Intelligence, IIT Madras (RBC-DSAI) for providing us with adequate resources. We also thank Gurneet Singh and Sarath Chandar for helping us in the data collection phase two and three respectively. Lastly, we thank all the AMT workers around the world and our in-house evaluators.

References

Antoine Bordes and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. *International Conference on Learning Representations*.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander H. Miller, Arthur Szlam, and Jason Weston. 2016. Evaluating prerequisite qualities for learning end-to-end dialog systems. *International Conference on Learning Representations*, abs/1511.06931.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pages 37–49.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model. *CoRR*, abs/1702.01932.

William H. Guss, James Bartlett, Phillip Kuznetsov, and Piyush Patil. 2017. Eigen: A step towards conversational ai. *Alexa Prize Proceedings*.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1766–1776.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014a. The second dialog state tracking challenge. In *Proceedings of the SIGDIAL 2014 Conference, The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 18-20 June 2014, Philadelphia, PA, USA*, pages 263–272.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014b. The third dialog state tracking challenge. In *2014 IEEE Spoken Language Technology Workshop, SLT 2014, South Lake Tahoe, NV, USA, December 7-10, 2014*, pages 324–329.

Ben Krause, Marco Damonte, Mihai Dobre, Daniel Duma, Joachim Fainberg, Federico Fancellu, Emmanuel Kahembwe, Jianpeng Cheng, and Bonnie L. Webber. 2017. Edina: Building an open domain socialbot with self-dialogues. *Alexa Prize Proceedings*.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2443–2453.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, pages 994–1003.

Yang Liu, Tim Paek, and Manasi Patwardhan, editors. 2018. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HTL 2018, New Orleans, Louisiana, USA, June 2-4, 2018, Demonstrations*. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, Laurent Charlin, and Joelle Pineau. 2015a. Incorporating unstructured textual knowledge sources into neural dialogue systems. In *Neural Information Processing Systems Workshop on Machine Learning for Spoken Language Understanding*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015b. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 285–294.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2017. Conversational AI: the science behind the alexa prize. *Alexa Prize Proceedings*.

Antoine Raux, Brian Langner, Dan Bohus, Alan W. Black, and Maxine Eskénazi. 2005. Let's go public! taking a spoken dialog system to the real world. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pages 885–888.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 172–180.

Lina Maria Rojas-Barahona, Milica Gasic, Nikola Mrksic, Pei-Hao Su, Stefan Ultes, Tsung-Hsien Wen, Steve J. Young, and David Vandyke. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 438–449.

Diane L. Schallert. 2002. Schema theory. *Literacy in America: An encyclopedia of history, theory, and practice Santa Barbara, CA*, pages 556–558.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083.

Stephanie Seneff, James R. Glass, David Goddeau, David Goodine, Lynette Hirschman, Hong C. Leung, Michael S. Phillips, Joseph Polifroni, and Victor Zue. 1991. Development and preliminary evaluation of the MIT ATIS system. In *Speech and Natural Language, Proceedings of a Workshop held at Pacific Grove, California, USA, February 19-22. 1991*.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. *International Conference on Learning Representations*.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *CoRR*, abs/1512.05742.

Iulian Vlad Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, Sai Mudumba, Alexandre de Brébisson, Jose Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, and Yoshua Bengio. 2017a. The octopus approach to the alexa competition: A deep ensemble-based socialbot. *Alexa Prize Proceedings*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 3776–3784.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017b. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3295–3301.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of*

Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers, pages 1577–1586.

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 *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 196–205.

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

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 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 189–198.

Jason D. Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *D&D*, 7(3):4–33.

Jason D. Williams, Antoine Raux, Deepak Ramachandran, and Alan W. Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference, The 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 22-24 August 2013, SUPELEC, Metz, France*, pages 404–413.