

Using a KG-Copy Network for Non-Goal Oriented Dialogues

Debanjan Chaudhuri^{1,2}, Md Rashad Al Hasan Rony^{1,2}, Simon Jordan³, and
Jens Lehmann^{1,2}

¹ Enterprise Information Systems Department, Fraunhofer IAIS, Dresden and
St. Augustin, Germany

{debanjan.chaudhuri, md.rashad.al.hasan.rony,
jens.lehmann}@iais.fraunhofer.de
<http://iais.fraunhofer.de>

² Smart Data Analytics Group, University of Bonn, Germany {chaudhur,
jens.lehmann}@cs.uni-bonn.de, s6mdrony@uni-bonn.de
<http://sda.tech>

³ Volkswagen Group Research, Wolfsburg, Germany
simon.jordan@volkswagen.de

Abstract. Non-goal oriented, generative dialogue systems lack the ability to generate answers with grounded facts. A knowledge graph can be considered an abstraction of the real world consisting of well-grounded facts. This paper addresses the problem of generating well-grounded responses by integrating knowledge graphs into the dialogue system’s response generation process, in an end-to-end manner. A dataset for non-goal oriented dialogues is proposed in this paper in the domain of soccer, conversing on different clubs and national teams along with a knowledge graph for each of these teams. A novel neural network architecture is also proposed as a baseline on this dataset, which can integrate knowledge graphs into the response generation process, producing well articulated, knowledge grounded responses. Empirical evidence suggests that the proposed model performs better than other state-of-the-art models for knowledge graph integrated dialogue systems.

Keywords: Non-goal oriented dialogues · knowledge grounded dialogues · knowledge graphs.

1 Introduction

With the recent advancements in neural network based techniques for language understanding and generation, there is an upheaved interest in having systems which are able to have articulate conversations with humans. Dialogue systems can generally be classified into goal and non-goal oriented systems, based on the nature of the conversation. The former category includes systems which are able to solve specific set of tasks for users within a particular domain, e.g. restaurant or flight booking. Non-goal oriented dialogue systems, on the other hand, are a first step towards chit-chat scenarios where humans engage in conversations

with bots over non-trivial topics. Both types of dialogue systems can benefit from added additional world knowledge [9],[11],[38].

For the case of non-goal oriented dialogues, the systems should be able to handle factoid as well as non-factoid queries like chit-chats or opinions on different subjects/domains. Generally, such systems are realized by using an extrinsic dialogue managers using intent detection subsequently followed by response generation (for the predicted intent) [1], [2]. Furthermore, in case of factoid queries posed to such systems, it is very important that they generate well articulated responses which are knowledge grounded. The systems must be able to generate a grammatically correct as well as factually grounded responses to such queries, while preserving co-references across the dialogue contexts. For better understanding, let us consider an example dialogue and the involved knowledge graph snippet in Figure 1. The conversation consists of chit-chat as well as factoid queries. For the factoid question "do you know what is the home ground of Arsenal?", the system must be able to answer with the correct entity (Emirates Stadium) along with a grammatically correct sentence; as well as handle co-references("its" in the third user utterance meaning the stadium). Ideally, for an end-to-end system for non-goal oriented dialogues, the system should be able to handle all these kind of queries using a single, end-to-end architecture.

There are existing conversation datasets supported by knowledge graphs for well-grounded response generation. [11] introduced an in-car dialogue dataset for multi-domain, task-oriented dialogues along with a knowledge graph which can be used to answer questions about the task the user wants to be assisted with. The dataset consists of dialogues from the following domains: calendar scheduling, weather information retrieval, and point-of interest navigation. For non-goal oriented dialogues, [10] proposed a dataset in the movie domain. The proposed dataset contains short dialogues for factoid question answering over movies or for recommendations. They also provide a knowledge graph consisting of triples as (s, r, o). Where s is the subject, r stands for relations and o being the object. An example of a triple from the dataset is: (Flags of Our Fathers, directed_by, Clint Eastwood). The movie dialogues can utilize this provided knowledge graph for recommendation and question answering purposes. However, this dataset only tackles the problem of factual response generation in dialogues, and not well articulated ones.

To cater to the problem of generating well articulated, knowledge grounded responses for non-goal oriented dialogue systems, we propose a new dataset in the domain of soccer. We also propose the KG-Copy network which is able to copy facts from the KGs in case of factoid questions while generating well-articulated sentences as well as implicitly handling chit-chats, opinions by generating responses like a traditional sequence-to-sequence model.

The contributions of the paper can be summarized as follows:

- A new dataset of 2,990 conversations for non-goal oriented dialogues in the domain of soccer, over various club and national teams.
- A soccer knowledge graph which consists of facts, as triples, curated from wikipedia.

	USER:	Hey what's up?	
	SYSTEM:	Nothing much, how are you? Do you want to talk soccer?	
	USER:	I am fine, thanks. Yeah, do you know what is the home ground of Arsenal ?	
	SYSTEM:	Arsenal's home ground is <i>Emirates Stadium</i> .	
	USER:	Oh ok, what's its capacity?	
	SYSTEM:	It has a capacity of 60,338.	

Subject	Predicate	Object
Arsenal	Home venue	Emirates Stadium
Emirates Stadium	capacity	60,338
Arsenal	chairman	Chips Keswick
Arsenal	Head coach	Unai Emery

Arsenal Knowledge Graph snippet.

Fig. 1. A conversation about the football club Arsenal and the Knowledge Graph involved.

- An end-to-end based, novel neural network architecture as a baseline approach on this dataset. The network is empirically evaluated against other state-of-the-art architectures for knowledge grounded dialogue systems. The evaluation is done based on both knowledge groundedness using entity-F1 score and also standard, automated metrics (BLEU) for evaluating dialogue systems.

The rest of the paper is organized as follows: we first introduce related work in Section 2. Then we cover the soccer dataset, which serves as background knowledge for our model in Section 3. The proposed model is explained in Section 4 and the training procedure is detailed in Section 5. In Section 6, we compare our model with other state-of-the-art models. We do a qualitative analysis of the model in Section 7, followed by an error analysis. In Section 8, finally we conclude.

2 Related Work

Systems that are able to converse with humans have been one of the main focus of research from the early days of artificial intelligence. Such conversational systems can be designed as generative or retrieval based. A system produces automatic responses from the training vocabulary for the former, while selecting a best response from a set of possible responses for the latter. Automatic

response generation was previously devised by [24] using a phrased-based generative method. Later onwards, sequence-to-sequence based neural network models has been mainly used for dialogue generation [19], [27], [29]. These models are further improved using hierarchical RNN based architectures for incorporating more contextual information in the response generation process [26]. Reinforcement learning-based end-to-end generative system were also proposed by [37] for jointly learning dialogue state-tracking [32] and policy learning [5].

[17] introduced the first multi-turn, retrieval based dataset which motivated a lot of further research on such systems. A lot of models are proposed on this dataset using both CNN [36], [3] and RNN [34], [31] based architectures. Both generative and retrieval based models can benefit from additional world knowledge as mentioned previously. However, the task of incorporating such additional knowledge (both structured and unstructured) into dialogue systems is challenging and is also a widely researched topic. [16], [35], [9] proposed architectures for incorporating unstructured knowledge into retrieval based systems. More recently, [12] incorporated unstructured knowledge as facts into generative dialogue systems as well.

Integration of structured knowledge comes in the form of incorporating knowledge graphs into the response generation process. [11] proposed a Key-Value retrieval network along with the in-car dataset (consisting of goal-oriented dialogues) for KG integration into sequence-to-sequence model. [20] proposed a generative model namely Mem2Seq for a task-oriented dialog system which combines multi-hop attention over memories with pointer networks. The model learns to generate dynamic queries to control the memory access. Mem2Seq is the current state-of-the-art on the in-car dataset. Further improvements on the task are proposed by [14] using joint embeddings and entity loss based regularization techniques. However, they learn the KG embeddings globally instead of per dialogue, so we evaluate our proposed system (KG-Copy network) against Mem2Seq.

Alongside the previously mentioned datasets for knowledge grounded dialogues, there is also a challenging dataset for complex sequential question answering which was introduced by [25]. It contains around 200K sequential queries that require a large KG to answer. The dataset contains questions that require inference and logical reasoning over the KG to answer. Although the dataset is the first non-goal oriented dataset which aims at knowledge graph integration, but it lacks proper conversational turns between utterances.

3 Soccer Dialogues Dataset

3.1 Wizard-of-Oz Style Data Collection

The proposed dataset for conversations over soccer is collected using AMT (Amazon Mechanical Turk) [8]. The dialogues are collected in an wizard-of-oz style [23] setup. In such a setup, humans believe they are interacting with machines, while the interaction is completely done by humans. The turkers, acting as users,

were instructed to initiate a conversation about the given team with any query or opinion or just have some small-talks. This initial utterance is again posted as another AMT task for replying, this time a different turker is acting as a system. Turkers assigned to the system role were asked to use Wikipedia to answer questions posed by the user. We encouraged the turkers to ask factual questions as well as posing opinions over the given teams, or have chit chat conversations. After a sequence of 7-8 utterances, the turkers were instructed to eventually end the conversation. A screenshot from the experimental setup is shown in Figure 2. We restricted the knowledge graph to a limited set of teams. The teams are picked based on popularity, the national teams chosen are: Sweden, Spain, Senegal, Portugal, Nigeria, Mexico, Italy, Iceland, Germany, France, Croatia, Colombia, Brazil, Belgium, Argentina, Uruguay and Switzerland. The club teams provided for conversing are: F.C. Barcelona, Real Madrid, Juventus F.C., Manchester United, Paris Saint Germain F.C., Liverpool F.C., Chelsea F.C., Atletico Madrid, F.C. Bayern Munich, F.C. Porto and Borussia Dortmund. We also encouraged people to converse about soccer without any particular team. The number of conversations are equally distributed across all teams. The statistics of the total number of conversations are given in Table 1.

Curious Football: Conversations over Football

Instructions: (Please read carefully)

- **Goal:** Based on the team/teams mentioned in the **cue** have a **question-answer** based **conversation** about **football (soccer)**.
- **Suggestions:** Ask questions about teams' performance, trivia, its players, and upcoming fixtures.
- **Answering:** Feel free to use **Wikipedia** or any other source for answering a questions.
- **Context:** Maintain a **context** of the conversations. Please ask questions or answer according to the last question in the **conversation box**.
- **Starting:** If conversation history is empty, start with a salutation.
- **Length:** A conversation length of 7-9 messages is optimum. Try to end a conversation after that.
- **Other:** We're **passionate** about football, and you might be too. But please keep the conversations **polite**.

Conversation History:
 Q1: Who is the best player in the world?
 A1: I would say Messi.

Type your question/answer here.

Fig. 2. AMT setup for getting conversations over soccer.

3.2 Ensuring Coherence

In order to ensure coherent dialogues between turkers, an additional task is created for each dialogue, where turkers were asked to annotate if the give dialogue is coherent or incoherent. Dialogues which are tagged incoherent by turkers are discarded.

Dataset	# of Dialogues	# of Utterances
Train	2,493	12,243
Validation	149	737
Test	348	1,727

Table 1. Statistics of Soccer Dataset.

3.3 Soccer Knowledge Graph

A KG in the context of this paper is a directed, multi-relational graph that represents entities as nodes, and their relations as edges, which can be used as an abstraction of the real world. KGs consists of triples of the form $(s,r,o) \in KG$, where s and o denote the subject and object entities, respectively, and r denotes their relation.

Following [6], we created a soccer knowledge graph from WikiData [30] which consists of information such as a team’s coach, captain and also information such as home ground and its capacity for soccer clubs. For information about players, we have parsed individual wikipedia pages of the teams and mined goals scored, position, caps, height and age of players. This ensures that the info in the KG is up to date. Finally, we curated the knowledge graphs for each team manually and added information such as jersey color. The KG schema is provided in 3 and additional statistics about KG and conversation is provided in table 2.

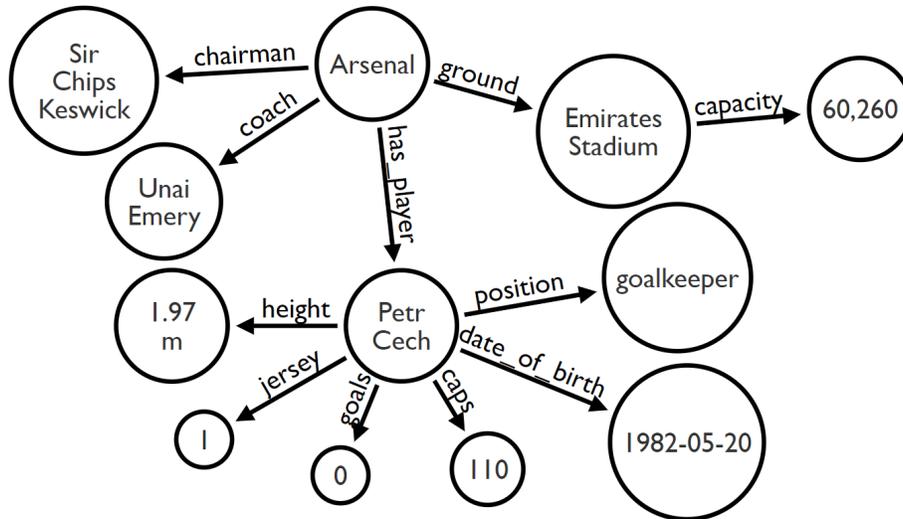


Fig. 3. Schema of the proposed Knowledge Graph for Arsenal.

Statistics	Count
Total Vocabulary Words (v)	4782
Avg. Number of Conversations/team	83
Avg. Number of Triples/team	148
Avg. Number of Entities/ team	108
Avg. Number of Relations/team	13

Table 2. KG statistics.

4 KG-Copy Model

The problem we are tackling in this paper is: given a knowledge graph (KG), and an input context in a dialogue, the model should be able to generate factual as well as well articulated response. During the dialogue generation process, at every time-step t , the model could either use the KG or generate a word from the vocabulary. We propose the KG Copy model which tackles this particular problem of producing well-grounded response generation.

KG-Copy is essentially a sequence-to-sequence encoder-decoder based neural network model, where the decoder can generate words either from the vocabulary or from the knowledge graph. The model is mainly influenced by the copynets approach [13]. However, unlike copynets, KG-Copy copies tokens from the local knowledge graph using a special gating mechanism. Here, local KG depicts the KG for the team the dialogue is about. We introduce the KG-Copy’s encoder, decoder and the gating mechanism below.

4.1 KG-Copy Encoder

The encoder is based on a recurrent neural network (RNN), more specifically a long-short term memory network (LSTM). It encodes the given input word sequence $X = [x_1, x_2, \dots, x_T]$ to a fixed length vector c . The hidden states are defined by

$$h_t = f_{enc}(x_t, h_{t-1}) \quad (1)$$

where f_{enc} is the encoder recurrent function and the context vector c is given by

$$c = \phi(h_1, h_2, \dots, h_T) \quad (2)$$

Here in, ϕ is the summarization function given the hidden states h_t . It can be computed by taking the last hidden state h_T or applying attention over the hidden states [4,18] and getting a weighted value of the hidden states (attention).

4.2 KG-Copy Decoder

The decoder is an attention based RNN (LSTM) model. The input to the decoder is the context c from the encoder along with h_T . At time-step t , the hidden-state

of the decoder is given by

$$h_t^d = f_{dec}(x_t, h_{t-1}) \quad (3)$$

Where f_{dec} is the recurrent function of the decoder. The decoder hidden-states are initialized using h_T and the first token is $\langle \text{sos} \rangle$. The attention mechanism [18]. The attention weights are calculated by concatenating the hidden states h_t^d along with h_t .

$$\alpha_t = \text{softmax}(W_s(\text{tanh}(W_c[h_t; h_t^d]))) \quad (4)$$

Here in, W_c and W_s are the weights of the attention model. The final weighted context representation is given by

$$\tilde{h}_t = \sum_t \alpha_t h_t \quad (5)$$

This representation is concatenated (represented by ;) with the hidden states of the decoder to generate an output from the vocabulary with size v .

The output is then given by

$$o_t = W_o([h_t; \tilde{h}_t^d]) \quad (6)$$

In the above equation, W_o are the output weights with dimension $\mathbf{R}^{h_{dim} \times v}$. h_{dim} is the dimension of the hidden layer of the decoder RNN.

4.3 Sentient Gating

The sentient gating, as mentioned previously, is inspired mainly by [13,21]. This gate acts as a sentinel mechanism which decides whether to copy from the local KG or to generate a word from training vocabulary (v). The final objective function can be written as the probability of predicting the next word during decoding based on the encoder hidden-states and the knowledge graph (KG).

$$p(y_t | h_t, h_1, KG) \quad (7)$$

The proposed gating is an embedding based model. At every time-step t , the input query and the input to the decoder are fed into the sentient gate. Firstly, a simple averaging of the input query embedding is done generating emb_q , which can be treated as an vector representation of the input context.

$$emb_q = \frac{1}{N} \sum (emb_{w_1} \dots emb_{w_t}) \quad (8)$$

emb_{w_t} is the embedding of the t^{th} word in the context. N.B. we only consider noun and verb phrases in the context to calculate emb_q . For the KG representation, an embedding average of the local KG's subject entity and relation labels for each triple is performed yielding a KG embedding emb_{kg} . We consider a total of k triples in the local KG.

Finally, the query embedding is matched with these KG embeddings using a similarity function (cosine similarity in this case).

$$kg_{sim} = \tanh(\cos(emb_q, emb_{kg}^1), \cos(emb_q, emb_{kg}^2) \dots \cos(emb_q, emb_{kg}^k)) \quad (9)$$

The input to the decoder at t is fed into the embedding too as mentioned previously yielding emb_d .

The final sentient value at t is given by :

$$s_t = \text{sigmoid}(W_{sent}[emb_q + emb_d; kg_{sim}; s_{t-1}]) \quad (10)$$

W_{sent} is another trainable parameter of the model and ";" is the concatenation operation. The final prediction is given by:

$$out_t = s_t * kg_{sim} + (1 - s_t) * o_t \quad (11)$$

The model is visualized in Figure 4.

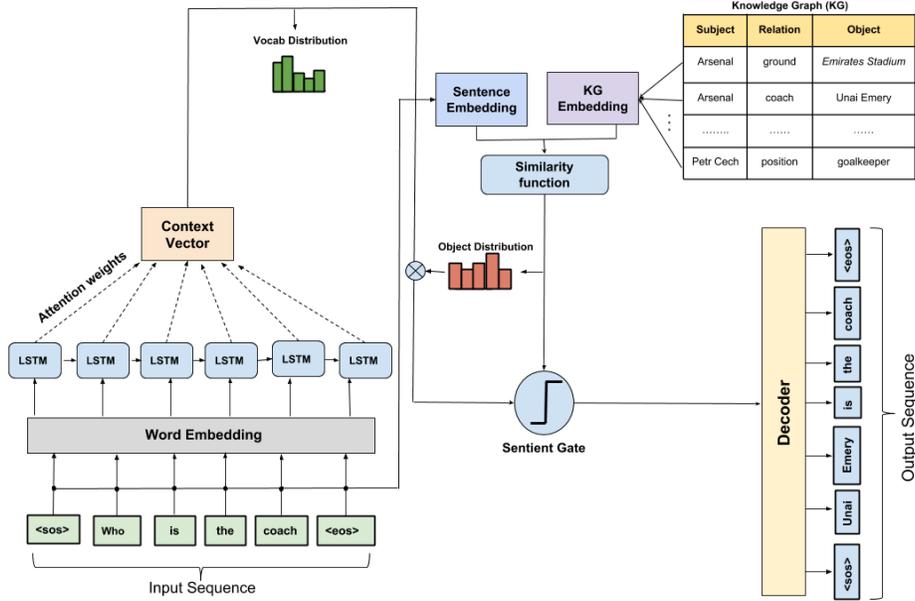


Fig. 4. KG-Copy Model Encoder-Decoder Architecture for Knowledge Grounded Response Generation.

5 Training and Model Hyper-parameters

5.1 Training Objective

The model is trained based on a multi-task objective, where the final objective is to optimize the cross-entropy based vocabulary loss (l_{vocab}) and also the binary

cross-entropy loss ($L_{sentient}$) for the sentient gate (s_g). This value is 1 if the generated token at that step comes from the KG, otherwise 0. For example, in the example provided in 1, for the 2nd system utterance, this value would be 1 for $t = 5$ (Emirates Stadium), but 0 for the previous time-steps. The total loss is given by:

$$L_{tot} = L_{vocab} + L_{sentient} \quad (12)$$

5.2 Training Details

To train the model, we perform a string similarity over KG for each of the questions in training data set to find which questions are answerable from the KG. Then we replace those answers with the position number of the triples where the answer (object) belongs in the KG, during pre-processing. This is followed by a manual step where we verify whether the input query is simple, factoid question or not and also the correctness of answer (object). The vocabulary is built only using the training data. No additional pre-processing is done for the validation and test sets except changing words to their corresponding indices in the vocabulary.

For training, a batch-size of 32 is used and the model is trained for 100 epochs. We save the model with the best validation f1-score and evaluate it on the test set. We apply Adam [15] for optimization with a learning rate of 1e-3 for the encoder and 5e-3 for the decoder. The size of the hidden layer of both the encoder and decoder LSTM is set to 64. We train the decoder RNN with teacher-forcing [33]. The input word embedding layer is of dimension 300 and initialized with pretrained fasttext [7] word embeddings. A dropout [28] of 0.3 is used for the encoder and decoder RNNs and 0.4 for the input embedding. The training process is conducted on a GPU with 3072 CUDA cores and a VRAM of 12GB. The soccer dataset (conversation and KG) and the KG-Copy model’s code are open-sourced ⁴ for ensuring reproducibility.

6 Evaluation

We compare our proposed model with Mem2Seq and a vanilla encoder-decoder with attention. We report the BLEU scores [22] and also the entity-F1 scores on both the proposed soccer dataset and the In-car dialogue dataset. The results show that our proposed model performs better than both the vanilla attention sequence-to-sequence models and Mem2Seq model across both metrics. Our model outperforms Mem2Seq by 1.51 in BLEU score and 15 % on entity-F1 score. It performs better than the vanilla sequence-to-sequence model by 1.21 on the BLEU metric on the soccer dataset. Interestingly, Mem2Seq performs better than the vanilla model on validation, but it fails to generalize on test set. The proposed model although has lower BLEU on the in-car dialogue dataset, but

⁴ https://github.com/SmartDataAnalytics/KG-Copy_Network

has a better entity f1 scores (by 19.4 %), implying stronger reasoning capabilities over entities and relations [20].

Model	BLEU		Entity-F1	
	Valid	Test	Valid	Test
Vanilla Encoder-decoder with Attention	1.04	0.82	–	–
Mem2Seq [20]	1.30	0.52	6.78	7.03
KG Copy (proposed model)	2.56	2.05	24.98	23.58

Table 3. Results on Soccer Dataset.

Model	BLEU	Entity-F1
Vanilla Encoder-decoder with Attention	8.4	10.3
Mem2Seq [20]	12.6	33.4
KG Copy (proposed model)	9.6	52.8

Table 4. Results on the In-car Dialogue Dataset.

7 Discussion

7.1 Qualitative Analysis

In this section, we will qualitatively analyze the response generation of our model along with the background knowledge integration (grounding) and compare it with both Mem2Seq and vanilla sequence-to-sequence models.

Some example response from test are given in Table 6. As seen, the KG-copy model is able to have more articulate responses compare to sequence-to-sequence and Mem2Seq models. The model is also able to form well articulate opinions compared to other models (2nd column)⁵.

Some more examples along with the response from our model are given in Table 7. As observed, all those responses are well grounded. The first response is factually correct and also a well-articulate one; interestingly, even the true human response on the other hand is not. The last response is knowledge ground but not well articulate. The model is also able to perform co-reference resolution implicitly while generating responses. To verify, let us consider another conversation between an user with the deployed KG-copy model.

⁵ Seq2Seq model has generated a more articulated response based on the given context but it is factually wrong: Senegal is nicknamed the Lions of Teranga and not Nigeria.

User utterance: i like the team pretty much
Response: i don't think they're a lot of winning.
User utterance: who is the captain of argentina ?
Response: lionel messi is the captain
User utterance: do you know the name of their coach ?
Response: lionel scaloni is the coach

In the last response, the model is able to identify that the pronoun "their" refers to the team and is able to maintain a knowledge grounded, as well as articulate responses even for relatively long dialogue turns. For time-step $t = 0$, the visualization of the sentient gating mechanism is provided in Figure 5. The vocabulary distribution is over v , and the object distribution here is over the local KG for the team.

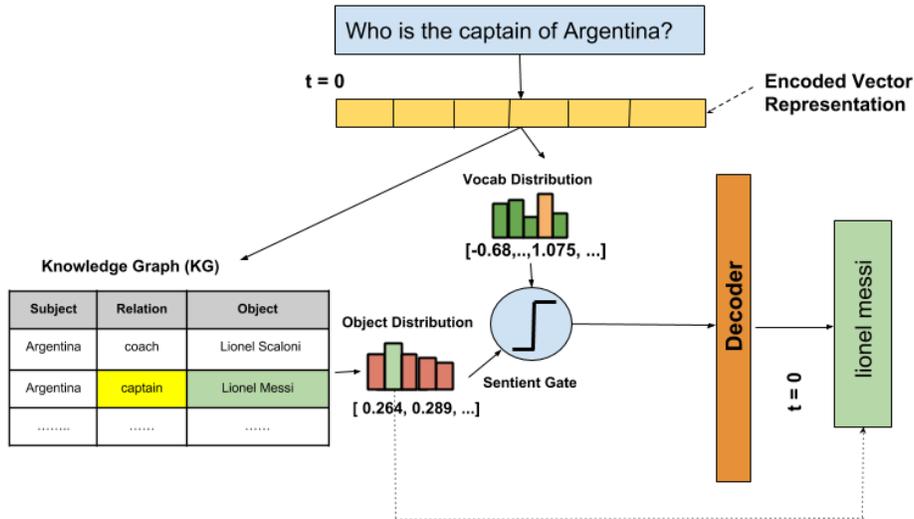


Fig. 5. Response Generation during Decoding for KG-Copy Model.

Furthermore, following [11], we did an internal survey regarding the responses generated by KG-Copy network, judging the quality of responses based on the context on a scale of 1-5 on correctness and human-like sentence formation. The former measures how correct the generated response is with respect to the true response from the turker, and the latter how grammatically correct the produced response is. We randomly pick 50 conversation utterances from the test set and report this human evaluation both on Mem2Seq and KG-copy in Table 5.

7.2 Error Analysis

Although the model is able to generate some well articulated, knowledge grounded responses for factual queries as seen in Tables 6 and 7, the model often fails in

Model	Correctness	Human-like
Mem2Seq	1.30	2.44
KG-Copy	2.26	3.88

Table 5. Human evaluation of generated responses.

Context Type	<i>factoid</i>	<i>opinions</i>
Input contexts	what is the name of the captain of mexico ?	I like this team.
True response	andres guardado (captain)	nigeria is a very well performing team and i like them a lot as well
Seq2Seq	The is is	They are nicknamed the Lions of Teranga.
Mem2Seq	Mexico is the	They are a
KG-Copy	andres guardado.	they are a good team.

Table 6. KG-copy Response for Factoid and non-Factoid queries.

producing factually correct responses as also evident from Table 5. More of those cases are analyzed below.

The model produces too generic and non-factual responses to queries about opinions about favorite players as shown in Table 8. This is mostly because the vocabulary size is relatively large compare to the size of training dialogues. This can be improved with more training data, especially with more knowledge grounded conversations. For the first response, the model is not able to interpret the question and generates a generic response. For the second case, the generated response is not factual because the question is about Argentina, but Eden Hazard is from a different team (Belgium).

The KG-copy model also often suffers when more complex quantitative and other reasoning skills are required to respond to the context. For example, for the first context in Table 9, the model needs to perform a count operation over the KG to answer it, which is currently unsupported. Similarly, for the second case the model would require better language inferencing to respond. The model also suffers from the problem of unknown words in the test set.

8 Conclusion and Future Work

In this paper, we introduce a new dataset for non-goal oriented, factual conversations over soccer (football). We also provide a knowledge graph for different club and national football teams which are the topic of these conversations.

Furthermore, we propose a relatively simple, novel, neural network architecture called KG-copy Network, as a baseline model, which can produce knowledge

input contexts	Turker Response	KG copy response
who is the captain of iceland?	aron gunnarsson	aron gunnarsson is the captain.
who is the captain of italy ?	chiellini	giorgio chiellini is the captain.
who is the coach for italy ?	i think roberto mancini	roberto mancini is the coach
who is the coach of bayern munich ?	niko kovac is the current manager	niko kovac

Table 7. KG Copy Model’s Knowledge Grounded Responses.

input contexts	True Response	KG copy response
who is senegal ’s best current player not including mane ?	keita balde diao	i think it is the best player in the world cup
who ’s your favorite player ?	messi	i think eden hazard is the best player

Table 8. Incorrect opinionated responses from KG-Copy model.

grounded responses as well as articulate responses via copying objects from the team KG based on the presented context of the question. Although the dataset is relatively small, the model can still learn the objective of producing grounded response as evident from the BLEU and entity-F1 scores compare to other models, and also from the examples provided in the paper. The proposed model also produces more knowledge grounded response (better entity f1 scores) on the in-car dialogue dataset [11] compared to other approaches. However, it should be noted that the BLEU scores in case of the non-goal oriented soccer dataset is lower compare to the goal oriented dataset (in-car). This can be attributed to the fact that the vocabulary size in case of the former is much larger (3 times), hence proving it to be a much harder problem. We also outlined weaknesses and limitations, e.g. for building factually correct responses, which can spur future research in this direction.

As a future work, we would like to consider a bigger study for gathering more knowledge-grounded, non-goal oriented conversations extending to more domains other than soccer. One of the problem with the dataset is that some responses from the turkers themselves are not articulate enough as evident from Table 7. To counter this, we would like to include more conversation verification steps and filter out conversations based on inter annotator agreements (IAA) between the turkers. Also, the proposed model can only respond to simple factoid questions based on word embedding based similarities between the context and the KG. We would like to extend the model to do better entity and relation linking between the query contexts and the knowledge graph in an end-to-end manner. The handling of out-of-vocabulary words also provides room for further

input contexts	how many world cups has the brazil team won ?	who was the top scorer in the world cup for belgium ?
True Response	brazil has won the fifa world cup five times	eden hazard
Predicted	they won the world cup	i think it was the top scorer for the world cup

Table 9. Incorrect factual responses from KG Copy model.

research. Moreover, we would also like to investigate recently proposed transformer or BERT based sequence-to-sequence models for the task of knowledge grounded response generation.

9 Acknowledgement

This work has been supported by the Fraunhofer-Cluster of Excellence “Cognitive Internet Technologies” (CCIT).

References

1. Agrawal, P., Suri, A., Menon, T.: A trustworthy, responsible and interpretable system to handle chit-chat in conversational bots. In: The Second AAAI Workshop on Reasoning and Learning for Human-Machine Dialogues (November 2018)
2. Akasaki, S., Kaji, N.: Chat detection in an intelligent assistant: Combining task-oriented and non-task-oriented spoken dialogue systems. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1308–1319. Association for Computational Linguistics, Vancouver, Canada (Jul 2017). <https://doi.org/10.18653/v1/P17-1120>
3. An, G., Shafiee, M., Shamsi, D.: Improving retrieval modeling using cross convolution networks and multi frequency word embedding. arXiv preprint arXiv:1802.05373 (2018)
4. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
5. Baird, L.: Residual algorithms: Reinforcement learning with function approximation. In: Machine Learning Proceedings 1995, pp. 30–37. Elsevier (1995)
6. Bergmann, T., Bunk, S., Eschrig, J., Hentschel, C., Knuth, M., Sack, H., SchÄijler, R.: In: Sabou, M., Blomqvist, E., Noia, T.D., Sack, H., Pellegrini, T. (eds.) I-SEMANTICS. pp. 146–149. ACM (2013)
7. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
8. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon’s mechanical turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science **6**(1), 3–5 (2011). <https://doi.org/10.1177/1745691610393980>, <https://doi.org/10.1177/1745691610393980>, pMID: 26162106

9. Chaudhuri, D., Kristiadi, A., Lehmann, J., Fischer, A.: Improving response selection in multi-turn dialogue systems by incorporating domain knowledge. In: Proceedings of the 22nd Conference on Computational Natural Language Learning. pp. 497–507. Association for Computational Linguistics, Brussels, Belgium (Oct 2018)
10. Dodge, J., Gane, A., Zhang, X., Bordes, A., Chopra, S., Miller, A., Szlam, A., Weston, J.: Evaluating prerequisite qualities for learning end-to-end dialog systems. ICLR (2016)
11. Eric, M., Krishnan, L., Charette, F., Manning, C.D.: Key-value retrieval networks for task-oriented dialogue. In: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. pp. 37–49. Association for Computational Linguistics (2017). <https://doi.org/10.18653/v1/W17-5506>
12. Ghazvininejad, M., Brockett, C., Chang, M.W., Dolan, B., Gao, J., Yih, W.t., Galley, M.: A knowledge-grounded neural conversation model. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
13. Gu, J., Lu, Z., Li, H., Li, V.O.: Incorporating copying mechanism in sequence-to-sequence learning. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1631–1640. Association for Computational Linguistics (2016). <https://doi.org/10.18653/v1/P16-1154>
14. Kassawat, F., Chaudhuri, D., Lehmann, J.: Incorporating joint embeddings into goal-oriented dialogues with multi-task learning. ESWC (2019)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Lowe, R., Pow, N., Serban, I., Charlin, L., Pineau, J.: Incorporating unstructured textual knowledge sources into neural dialogue systems. In: Neural Information Processing Systems Workshop on Machine Learning for Spoken Language Understanding (2015)
17. Lowe, R., Pow, N., Serban, I., Pineau, J.: The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. arXiv preprint arXiv:1506.08909 (2015)
18. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
19. Luong, T., Sutskever, I., Le, Q., Vinyals, O., Zaremba, W.: Addressing the rare word problem in neural machine translation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 11–19. Association for Computational Linguistics, Beijing, China (Jul 2015). <https://doi.org/10.3115/v1/P15-1002>
20. Madotto, A., Wu, C.S., Fung, P.: Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1468–1478. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)
21. Merity, S., Xiong, C., Bradbury, J., Socher, R.: Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843 (2016)
22. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)

23. Rieser, V., Lemon, O.: Learning effective multimodal dialogue strategies from wizard-of-oz data: Bootstrapping and evaluation. *Proceedings of ACL-08: HLT* pp. 638–646 (2008)
24. Ritter, A., Cherry, C., Dolan, W.B.: Data-driven response generation in social media. In: *Proceedings of the conference on empirical methods in natural language processing*. pp. 583–593. Association for Computational Linguistics (2011)
25. Saha, A., Pahuja, V., Khapra, M.M., Sankaranarayanan, K., Chandar, S.: Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In: *Thirty-Second AAAI Conference on Artificial Intelligence* (2018)
26. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: *Thirtieth AAAI Conference on Artificial Intelligence* (2016)
27. Shang, L., Lu, Z., Li, H.: Neural responding machine for short-text conversation. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Conference on Natural Language Processing (Volume 1: Long Papers)*. pp. 1577–1586. Association for Computational Linguistics, Beijing, China (Jul 2015). <https://doi.org/10.3115/v1/P15-1152>
28. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **15**(1), 1929–1958 (2014)
29. Vinyals, O., Le, Q.V.: A neural conversational model. *ArXiv* **abs/1506.05869** (2015)
30. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledgebase. *Commun. ACM* **57**(10), 78–85 (Sep 2014). <https://doi.org/10.1145/2629489>, <http://doi.acm.org/10.1145/2629489>
31. Wang, S., Jiang, J.: Learning natural language inference with lstm. In: *Proceedings of NAACL-HLT 2016* (2016)
32. Williams, J., Raux, A., Ramachandran, D., Black, A.: The dialog state tracking challenge. In: *Proceedings of the SIGDIAL 2013 Conference*. pp. 404–413 (2013)
33. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. *Neural computation* **1**(2), 270–280 (1989)
34. Wu, Y., Wu, W., Xing, C., Zhou, M., Li, Z.: Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. vol. 1, pp. 496–505 (2017)
35. Xu, Z., Liu, B., Wang, B., Sun, C., Wang, X.: Incorporating loose-structured knowledge into lstm with recall gate for conversation modeling. *arXiv preprint arXiv:1605.05110* **3** (2016)
36. Yan, R., Song, Y., Wu, H.: Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. pp. 55–64. ACM (2016)
37. Zhao, T., Eskenazi, M.: Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. *arXiv preprint arXiv:1606.02560* (2016)
38. Zhu, W., Mo, K., Zhang, Y., Zhu, Z., Peng, X., Yang, Q.: Flexible end-to-end dialogue system for knowledge grounded conversation. *arXiv preprint arXiv:1709.04264* (2017)