Knowledge Base Adaptation for Task Oriented Dialog

Thesis submitted by

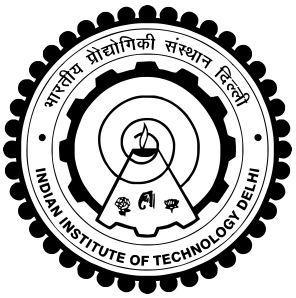
Nikhil Gupta
2014CS50462

under the guidance of

Prof. Mausam

in partial fulfilment of the requirements
for the award of the degree of

Bachelor and Master of Technology



Department Of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY DELHI

June 2019

THESIS CERTIFICATE

This is to certify that the thesis titled **Knowledge Base Adaptation for Task Oriented Dialog**, submitted by **Nikhil Gupta**, to the Indian Institute of Technology, Delhi, for the award of the degree of **Bachelor and Master of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Mausam
Professor
Dept. of Computer Science
IIT-Delhi, 600 036

Place: New Delhi

Date: 28th June 2019

ACKNOWLEDGEMENTS

I would like to extend thanks to Microsoft for graciously providing me a VM to work on. I would thank my advisor Mausam for his constant support and insights on this project. I would like to thank Dinesh for his suport and effort and guiding me along the way for this journey.

We thank Danish Contractor, Gaurav Pandey and Sachindra Joshi for their comments on an earlier version of this work. This work is supported by IBM AI Horizons Network grant, an IBM SUR award, grants by Google, Bloomberg and 1MG, and a Visvesvaraya faculty award by Govt. of India. We thank Microsoft Azure sponsorships, and the IIT Delhi HPC facility for computational resources.

ABSTRACT

KEYWORDS: LaTeX; Thesis; Style files; Format.

A LaTeX class along with a simple template thesis are provided here. These can be used to easily write a thesis suitable for submission at IIT-Delhi. The class provides options to format PhD, MS, M.Tech. and B.Tech. thesis. It also allows one to write a synopsis using the same class file. Also provided is a BIBTeX style file that formats all bibliography entries as per the IITD format.

The formatting is as (as far as the author is aware) per the current institute guidelines.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Task-oriented dialog agents converse with a user with the goal of accomplishing a specific task and often interact with a knowledge-base (KB). For example, a restaurant reservation agent [6] will be grounded to a KB that contains the names of restaurants, and their details.

In real-world applications, the KB information could change over time. For example, (1) a KB associated with a movie ticket booking system gets updated every week based on new film releases, and (2) a restaurant reservation agent, trained with the knowledge of eateries in one city, may be deployed in other cities with an entirely different range of establishments. In such situations, the system should have the ability to conform to new-found knowledge unseen during its training. Ideally, the training algorithm must learn to disentangle the language model from the knowledge interface model. This separation will enable the system to generalize to KB modifications, without a loss in performance.

Moreover, for achieving good progress towards the user's task, the agent must also retain the ability to draw inferences based on past utterances and the KB. Notably, we find that existing approaches either achieve this disentanglement or effective progress towards the task, but not both.

For instance, Mem2Seq [8] exhibits satisfactory performance when tested on the training KB. It represents the dialog history and the KB knowledge as a *bag of words* in a flat memory arrangement. This enables Mem2Seq to revisit each word several times, as needed, obtaining good performance. But at the same time, flat memory prevents it from capturing any surrounding context – this deteriorates its performance rapidly when the amount of new unseen information in the KB increases, as shown in Figure **??**. On the other hand, the performance of copy augmented sequence-to-sequence network (Seq2Seq+Copy) [3], is robust to changes in the KB, but fails to achieve acceptable task-oriented performance. It captures context by representing the entire dialog history as one continuous *sequence*. However, it can be difficult for a sequence encoder to reason over long dialogs found in real-world datasets and its ability to learn the task gets hampered.

We propose BoSsNet, a novel network that effectively disentangles the language and knowledge models, and also achieves state-of-the-art performance on three existing datasets.

To achieve this, BoSsNet makes two design choices. First, it encodes the conversational input as a *bag of sequences* (BoSs) memory, in which the input representation is built at two levels of abstraction. The *higher level* flat memory encodes the KB tuples and utterances to facilitate effective inferencing over them. The *lower level* encoding of each individual

utterance and tuple is constructed via a sequence encoder (Bi-GRU). This enables the model to maintain the sequential context surrounding each token, aiding in better interpretation of unseen tokens at test time. Second, we augment the standard cross-entropy loss used in dialog systems with an additional loss term to encourage the model to only copy KB tokens in a response, instead of generating them via the language model. This combination of sequence encoding and additional loss (along with dropout) helps in effective disentangling between language and knowledge.

We perform evaluations over three datasets – bAbI [1], CamRest [14], and Stanford Multi-Domain Dataset [2]. Of these, the last two are real-world datasets. We find that BoSsNet is competitive or significantly better on standard metrics in all datasets as compared to state-of-the-art baselines. We also introduce a *knowledge adaptability* (KA) evaluation, in which we systematically increase the percentage of previously unseen entities in the KB. We find that BoSsNet is highly robust across all percentage levels. Finally, we also report a human-based evaluation and find that BoSsNet responses are frequently rated higher than other baselines.

Overall, our contributions are:

1. We propose BoSsNet, a novel architecture to disentangle the language model from knowledge incorporation in task-oriented dialogs.
2. We introduce a *knowledge adaptability* evaluation to measure the ability of dialog systems to scale performance to unseen KB entities.
3. Our experiments show that BoSsNet is competitive or significantly better, measured via standard metrics, than the existing baselines on three datasets.

We release our code and *knowledge adaptability* (KA) test sets for further use by the research community. `https://github.com/dair-iitd/BossNet`.

Chapter 2

BACKGROUND

Compared to the traditional slot-filling based dialog [16, 13, 15], end-to-end training methods (e.g., [1], this work) do not require handcrafted state representations and their corresponding annotations in each dialog. Thus, they can easily be adapted to a new domain. We discuss end-to-end approaches along two verticals: 1) decoder: whether the response is retrieved or generated and 2) encoder: how the dialog history and KB tuples are encoded.

Most of the existing end-to-end approaches *retrieve* a response from a pre-defined set [1, 7, 11]. These methods are generally successful when they have to provide boilerplate responses – they cannot construct responses by using words in KB not seen during training. Alternatively, generative approaches are used where the response is *generated* one word at a time [3, 8]. These approaches mitigate the unseen entity problem by incorporating the ability to copy words from the input [12, 4]. The copy mechanism has also found success in summarization [9, 10] and machine translation [5]. BoSsNet is also a copy incorporated generative approach.

For encoding, some approaches represent the dialog history as a sequence [3, 5]. Unfortunately, using a single long sequence for encoding also enforces an order over the set of KB tuples making it harder to perform inferencing over them. Other approaches represent the dialog context as a bag. Original Memory Networks [1] and its extensions encode each memory element (utterance) as an average of all constituent words – this cannot point to individual words, and hence cannot be used with a copy mechanism. Mem2Seq encodes each word individually in a flat memory. Unfortunately, this loses the contextual information around a word, which is needed to decipher an unseen word. In contrast, BoSsNet uses a bag of sequences encoding, where KB tuples are a set for easier inference, and also each utterance is a sequence for effectively learning when to copy.

Chapter 3

APPROACH

Chapter 4

WORK DIVISION

Chapter 5

IMPLEMENTATION

Chapter 6

RESULTS

6.1 Datasets

We perform experiments on three task-oriented dialog datasets: bAbI Dialog [1], CamRest [14], and Stanford Multi-Domain Dataset [2].

bAbI Dialog consists of synthetically generated dialogs with the goal of restaurant reservation. The dataset consists of five different tasks, all grounded to a KB. This KB is split into two mutually exclusive halves. One half is used to generate the train, validation, and test sets, while the other half is used to create a second test set called the OOV test set.

CamRest is a human-human dialog dataset, collected using the Wiz-of-Oz framework, also aimed at restaurant reservation. It is typically used to evaluate traditional slot filling systems. In order to make it suitable for end-to-end learning, we stripped the handcrafted state representations and annotations in each dialog, and divided the 676 available dialogs into train, validation, and test sets (406, 135, and 135 dialogs, respectively).

Stanford Multi-Domain Dataset (SMD) is another human-human dialog dataset collected using the Wiz-of-Oz framework. Each conversation is between a driver and an in-car assistant. The other datasets consist of dialogs from just one domain (restaurant reservation), whereas SMD consists of dialogs from multiple domains (calendar scheduling, weather information retrieval, and navigation).

6.1.1 Knowledge Adaptability (KA) Test Sets

Each bAbI dialog task has an additional OOV test set, which helps to evaluate a model's robustness to change in information in the KB. A model that perfectly disentangles language and knowledge should have no drop in accuracy on the OOV test set when compared to the non-OOV test set. To measure the degree of disentanglement in a model, we generated 10 additional test sets for each real-world corpus by varying the percentage (in multiples of 10) of unseen entities in the KB. We systematically picked random KB entities and replaced all their occurrences in the dialog with new entity names. We will refer to these generated dialogs as the *Knowledge Adaptability* (KA) test sets.

6.2 Baselines

We compare BOSSNET against several existing end-to-end task-oriented dialog systems. These include retrieval models, such as the query reduction network (QRN) [11], memory network (MN) [1], and gated memory network (GMN) [7].

We also compare against generative models such as a sequence-to-sequence model (Seq2Seq), a copy augmented Seq2Seq (Seq2Seq+Copy) [5], and Mem2Seq [8].[1] For fairness across models, we do not compare against key-value retrieval networks [2] as they simplify the dataset by canonicalizing all KB words in dialogs.

We noticed that the reported results in the Mem2Seq paper are not directly comparable, as they pre-processed[2] training data in SMD and bAbI datasets. For fair comparisons, we re-run Mem2Seq on the original training datasets. For completeness we mention their reported results (with pre-processing) as Mem2Seq*.

6.3 Evaluation Metrics

We evaluate BOSSNET and other models based on their ability to generate valid responses. The per-response accuracy [1] is the percentage of generated responses that exactly match their respective gold response. The per-dialog accuracy is the percentage of dialogs with all correctly generated responses. These accuracy metrics are a good measure for evaluating datasets with boilerplate responses such as bAbI.

To quantify performance on other datasets, we use BLEU [**?**] and Entity F1 [3] scores. BLEU measures the overlap of n-grams between the generated response and its gold response and has become a popular measure to compare task-oriented dialog systems. Entity F1 is computed by micro-F1 over KB entities in the entire set of gold responses.

6.4 Human Evaluation

We use two human evaluation experiments to compare (1) the *usefulness* of a generated response with respect to solving the given task, and (2) the *grammatical correctness* and *fluency* of the responses on a 0–3 scale. We obtain human annotations by creating Human Intelligence Tasks (HITs) on Amazon Mechanical Turk (AMT). For each test condition

[1] We thank the authors for releasing a working code at `https://github.com/HLTCHKUST/Mem2Seq`

[2] Mem2Seq used the following pre-processing on the data: 1) The subject (restaurant name) and object (rating) positions of the rating KB tuples in bAbI dialogs are flipped 2) An extra fact was added to the navigation tasks in SMD which included all the properties (distance, address, etc.) combined together as the subject and *poi* as the object. See Appendix.

(percentage of unseen entities), we sampled 50 dialogs from Camrest and SMD each, and two AMT workers labeled each system response for both experiments, resulting in 200 labels per condition per dataset per system. We evaluate four systems in this study, leading to a total of 1600 labels per condition. The detailed setup is given in the Appendix.

6.5 Training

We train BoSsNet using an Adam optimizer [**?**] and apply gradient clipping with a clip-value of 40. We identify hyper-parameters based on the evaluation of the held-out validation sets. We sample word embedding, hidden layer, and cell sizes from {64, 128, 256} and learning rates from $\{10^{-3}, 5{\times}10^{-4}, 10^{-4}\}$. The hyper-parameter γ in the loss function is chosen between [0-1.5]. The Disentangle Label Dropout rate is sampled from {0.1, 0.2}. The number of hops for multi-hop attention in the encoder is sampled from {1, 3, 6}. The best hyper-parameter setting for each dataset is reported in the Appendix.

Appendix A

CODE SNIPPETS

REFERENCES

[1] A. Bordes and J. Weston. Learning end-to-end goal-oriented dialog. In *International Conference on Learning Representations*, 2017.

[2] M. Eric, L. Krishnan, F. Charette, and C. D. Manning. Key-value retrieval networks for task-oriented dialogue. In *Dialog System Technology Challenges, Saarbrücken, Germany, August 15-17, 2017*, pages 37–49, 2017.

[3] M. Eric and C. D. Manning. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. *arXiv preprint arXiv:1701.04024*, 2017.

[4] J. Gu, Z. Lu, H. Li, and V. O. Li. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*, 2016.

[5] C. Gulcehre, S. Ahn, R. Nallapati, B. Zhou, and Y. Bengio. Pointing the unknown words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 140–149. Association for Computational Linguistics, 2016.

[6] M. Henderson, B. Thomson, and S. Young. Word-based dialog state tracking with recurrent neural networks. In *In Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299, 2014.

[7] F. Liu and J. Perez. Gated end-to-end memory networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 1–10, 2017.

[8] A. Madotto, C. Wu, and P. Fung. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2018.

[9] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290. Association for Computational Linguistics, 2016.

[10] A. See, P. J. Liu, and C. D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083. Association for Computational Linguistics, 2017.

[11] M. Seo, S. Min, A. Farhadi, and H. Hajishirzi. Query-reduction networks for question answering. In *International Conference on Learning Representations*, 2017.

[12] O. Vinyals, M. Fortunato, and N. Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700, 2015.

[13] T. Wen, D. Vandyke, N. Mrkšíc, M. Gašíc, L. Rojas-Barahona, P. Su, S. Ultes, and S. Young. A network-based end-to-end trainable task-oriented dialogue system. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017-Proceedings of Conference*, volume 1, pages 438–449, 2017.

[14] T.-H. Wen, M. Gasic, N. Mrkšić, L. M. Rojas Barahona, P.-H. Su, S. Ultes, D. Vandyke, and S. Young. Conditional generation and snapshot learning in neural dialogue systems. In *EMNLP*, pages 2153–2162, Austin, Texas, November 2016. ACL.

[15] J. D. Williams, K. Asadi, and G. Zweig. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 665–677, 2017.

[16] J. D. Williams and S. Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422, 2007.

CITATIONS

1. Yavuz et Al. **DEEPCOPY: Grounded Response Generation with Hierarchical Pointer Networks** *NIPS* (2018).

2. Ebrahimi et Al. **Reasoning over RDF Knowledge Bases using Deep Learning** *arXiv* (2018).

3. Singh et Al. **Towards VQA Models That Can Read** *CVPR* (2019).

4. Golchha et Al. **Courteously Yours: Inducing courteous behavior in Customer Care responses using Reinforced Pointer Generator Network** *NAACL-HLT* (2019).