Augmenting Conversational Dialogue Datasets with Commonsense and Adaptive Local Knowledge

CSCI 662: Advanced Natural Language Processing
Fall 2020 Final Report

Hyundong Cho and **Basel Shbita**
University of Southern California
{jcho, shbita}@isi.edu

Abstract

Modern chit-chat dialogue systems are based on powerful language models trained with massive dialogue data, and a recent line of work focuses on incorporating external knowledge sources to ground these models with commonsense. However, most use a static knowledge source that is not updated with information introduced from an ongoing conversation or only work with single-turn dialogue pairs, which is not truly 'dialogue'. We present a dialogue dataset augmentation framework and expand the multi-turn Persona Chat dataset with a turn-level adaptive local knowledge base that maintains the speaker's persona and knowledge relevant to the current conversation. We evaluate the effectiveness of this approach with fine-tuned language models on generating coherent responses in the multi-turn dialogue setting and identify limitations that must be tackled.

1 Introduction

Recent works on massive data-driven language models provide evidence that capture linguistic patterns that allow them to perform well on many NLP tasks, including open-domain dialogue (Devlin et al., 2018; Liu et al., 2019; Radford et al., 2019; Yang et al., 2019; Wolf et al., 2019; Zhang et al., 2020b). While these data-driven models are impressive, they have no mechanism to talk about knowledge that is not in the scopes of the data that they were trained on. Filling such gaps requires more complex reasoning for which the language models need to be enriched with suitable background knowledge.

In such a setting, sharing the same understanding of the world with its constituents are important for effective interactions. This understanding is not explicitly established for every conversation. They are implicit and assumed to be known. However, this understanding need not be stagnant either. An ideal agent can also contribute and adapt to the reality established throughout the conversation.

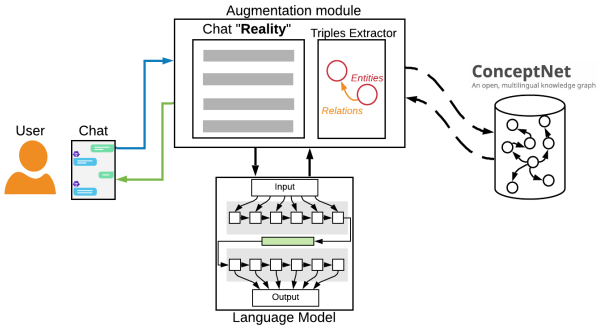

Figure 1: A high-level illustration of **AALK** and the subsequent training with a language model.

The main limitation of previous works that incorporate external knowledge to dialogue generation is their reliance on a static knowledge graph or knowledge base that remains unchanged throughout the entire conversation (Young et al., 2017; Parthasarathi and Pineau, 2018; Moon et al., 2019; Zhou et al., 2018; Zhang et al., 2020a). This inflexibility can make a conversation predictable and uninteresting, which is detrimental to open-domain dialogue systems. For creative dialogue, such as those taking place in fictional settings or improvisational theatre, commonsense can set the basis but the reality can be shaped as the conversation takes place.

To this end, we propose a dialogue dataset augmentation process called *Augment with Adaptive Local Knowledge* (**AALK**) that adds relevant commonsense knowledge and maintains an adaptive user persona. Together, they represent local knowledge about the agent's understanding of the world and its persona. Specifically, we use a set of information extraction modules and heuristics for extracting relevant information from DBpedia[1] and ConceptNet[2] to expand Persona Chat by updating

[1] https://wiki.dbpedia.org/
[2] https://concpetnet.io

local knowledge and agent profile at the turn-level (Speer et al., 2017; Zhang et al., 2018). Our preliminary results with a dialogue model trained on the augmented show no noticeable improvements due to misalignment in injected information and target response. We identify rooms for improvement that can tackle these shortcomings.

2 Problem Definition

We formalize our knowledge-grounded dialogue problem as the following: Given (i) a a set of dialogue turns $X = \{X_1, X_2, ..., X_t\}$; (ii) a commonsense knowledge graph $G = \{\tau_1, \tau_2, ..., \tau_N\}$, where each $\tau = (h, r, t)$ is a triple (head entity, relation, tail entity); and (iii) a system's knowledge $\Pi = \{\pi_1, \pi_2, ..., \pi_{\tilde{D}}\}$, where each π is unstructured natural text describing what the system knows about the current conversation, the goal is to generate a proper response Y. It is important to note that Π is not a constant, due to the dynamic nature of the continuously updated profile of the user. More specifically, we model the probability $P(Y_t \mid G, \Pi_t, X_t, X_{t-1}, ..., X_0)$, where t represents the current turn.

3 Approach and Implementation

We call our augmentation process *Augment with Adaptive Local Knowledge* (**AALK**). We present **AALK** and the subsequent model training with Persona Chat as an example.

Data augmentation **AALK** involves two main steps for every utterance: (i) extract relevant commonsense knowledge from ConceptNet or DBpedia and add it to the local knowledge base and (ii) identify profile statements and add it to the user profile. A high-level illustration of our approach is shown in Figure 1. For Persona Chat, we leverage the initial set of profile information and apply step (i) to it as well, treating each profile description as an utterance to add information for.

First, for each utterance, we identify the relevant triples: the pairs of concepts/entities connected by a relation. We extract the relevant data from the given text using linguistic and statistical processing with a dedicated pipeline we developed with *SpaCy* (Honnibal and Montani, 2017). The extraction of the entities was implemented with heuristics driven by the parts of speech (POS) annotation tags and the dependency tree of the sentence. For the extraction of the relation, we used rule-based matching with the hypothesis that the relation (i.e. predicate) is the main verb in a sentence. The relation is compounded with auxiliary, preposition and negation words, if present.

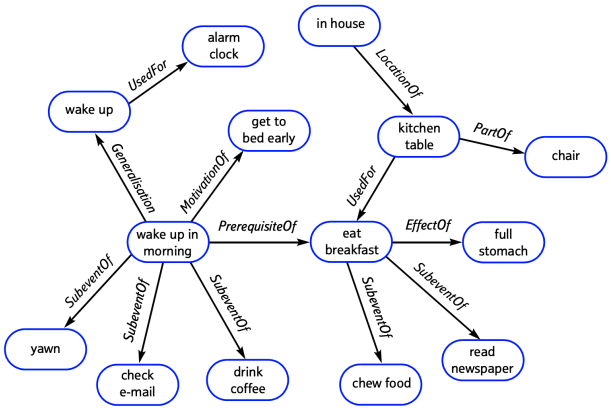

Figure 2: A subset of ConceptNet's semantic network

Next, once we constructed the relevant triples for each user utterance, we retrieve a piece of common sense knowledge from ConceptNet based on a 'focus' term in each extracted triple. This term, which could be either a relation or a concept, is chosen based on rules, such as ignoring common verbs like `do, think` and pronouns such as `I, you`, which are either too general or not an entry in ConceptNet.

Then, we retrieve a ranked list of commonsense statements from the KG. ConceptNet uses a closed set of selected relations intended to represent the relationships between concepts, as seen in figure 2. The chosen statement is determined by two factors: (i) its crowd-sourced weight in ConceptNet and (ii) the predetermined relations hierarchy that we developed. In this hierarchy, we prefer certain relations that we consider more interesting than others (e.g., the relation `CausesDesire` is preferred over `LocatedNear`).

The 'Augmented Knowledge' column in table 1 shows examples of profile statements we retrieved for each dialogue. If an entity with a special type (e.g., `PERSON`, `ORG`, `EVENT`, etc...) is detected, we use ConceptNet's external links to *DBpedia* (Auer et al., 2007) to retrieve data about the entity. An example for such an instance can be seen in the second row in the table (*'Taylor Alison Swift is an American singer-songwriter'*).

Lastly, we construct the agent profile to be added to the original persona descriptions based on the triples we extracted. This enables establishing a ground that is not necessarily based on commonsense but may push the conversation to be more engaging for the user. An example of such can be

Dialogue	Extracted Triples	Augmented Knowledge	User-Profile Knowledge
User: i do not like crowds Bot: working out is a great way to burn off steam . do you like country music ? User: a little bit . i can get into taylor swift .	(i, do not like, crowds) (working out, is, great steam) (you, like, country music) (i, can get, taylor swift)	[*'You are likely to find people in crowds'*, *'country music is a type of folk music'*, *'Taylor Alison Swift is an American singer-songwriter'*]	[*'i do not like crowds'*, *'i can get taylor swift'*]
User: i make time stop . i've a superpower . i'm a super hero . Bot: that is really cool that you can do that User: i love living in the clouds the best . do you have powers ? Bot: no i don't , i faint when i see blood User: i don't like blood . i vomit .	(i, have, superpower) (i, love, clouds) (i, don't like, blood)	[*'You are likely to find clouds in the sky'*, *'Something you find at a hospital is blood'*]	[*'i have superpower'*, *'i love clouds'*, *'i don't like blood'*]

Table 1: Example dialogues and their corresponding extracted and retrieved attributes and data

seen in the third row in the table, where the fact *'i have superpower'* about the user is added to the agent's profile. We show sample results of **AALK** in Table 1.

Model We evaluate the usefulness of **AALK** by training a relatively simple dialogue model. With the augmented Persona Chat, we train a *TransferTransfo* model (Wolf et al., 2019) that applies transfer learning to transformer models (Vaswani et al., 2017). This model is trained on the joint loss of the standard language modeling task and the next sentence prediction task. Instead of fine-tuning the GPT model, we use GPT-2 (Radford et al., 2019). Augmented knowledge and persona are treated equally in that they are both prepended to the dialogue history. The dialogue history is separated by turn with special tokens that indicate whose turn it is to respond. For each training sample, only the language modeling loss for the last turn is calculated against the target response.

4 Experiments Details and Results

Training We train with only the validation set and its augmented version as a proof-of-concept experiment to test whether the additional information and explicit addition to the persona of the extracted profile statement enhances the model's capacity to generate knowledge-grounded responses in a multi-turn setting and accommodate dynamically introduced new persona information. 10% of the validation set is held out as the development set, such that we train with 6,240 training samples.

We train two *TransferTransfo* models, one on the original dataset and the other on the augmented version, for four epochs using a batch size of 4 and learning rate of $6.25e^{-5}$ on an Adam optimizer (Kingma and Ba, 2014). We consider up to four previous turns in predicting the next response and consider two other distractor candidates for the next sentence prediction task. The language modeling loss is weighted twice as much as the next sentence prediction loss. Other training configurations that are not mentioned are identical to the default settings of the *TransferTransfo* model. The models are trained on 2 V100 GPUs and take about 40 minutes each in total.

	LM Loss	*Accuracy*	*Perplexity*
Persona Chat	3.01	**0.49**	**20.30**
+ **AALK**	3.01	0.47	20.37

Table 2: Comparison of a *TransferTransfo* model trained on Persona Chat and another trained on the version augmented with **AALK**.

Evaluation Open-domain dialogue is a high-entropy task where many responses can be suitable for a given dialogue context, and therefore using metrics such as BLEU, ROUGE, or METEOR (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005), which measures overlap with a reference text, is inadequate. However, we still rely on automatic evaluation to get a sense of relative performance before moving on to human evaluation.

For automatic evaluation, we follow the work by Wolf et al. (2019) and use language modeling loss, accuracy on the next sentence prediction task, and perplexity. The comparison results are summarized in Table 2. On all metrics, the model performed worse when trained and validated on the augmented data. These results indicate that we should reexamine **AALK** before proceeding to human evaluations.

5 Discussion

The poor results do not necessarily mean that **AALK** is an inadequate mechanism for augmenting dialogue data with external knowledge and adaptive user profile information. Upon inspection, we see that there is not a lot of dialogue in Persona

Chat that reflects our objective of incorporating commonsense and local knowledge. Even when useful or interesting information is augmented by **AALK**, it is often not useful for generating the target response. We hypothesize that this disconnect makes the augmented text act as noise, doing more harm than good. In order for **AALK** to be more effective, we will need a quality assurance step that verifies the usefulness of the information that is to be augmented in relation to any of the future responses. This may be achieved with a natural language inference model that indirectly measures the relevance of the added information and the target response.

Apart from the misalignment with augmented information and the target response, **AALK** itself has several limitations. First, the triples extraction pipeline we constructed is a bottleneck to effectively extracting knowledge from generated responses. When applied to complicated information that is split along sentences, our pipeline fails to capture useful data from the user. **AALK**'s effectiveness diminishes quickly as errors in the early stages can lead to extracting inaccurate profile information or irrelevant knowledge from external knowledge resources. We need a better understanding of failure cases and develop approaches for such cases that are not easily handled using dependency trees. Alternatively, an approach that bypasses manual heuristics and modular components for a more end-to-end framework will be more ideal.

Also, KGs can be hard to traverse to find related concepts if they are not directly connected or if they require disambiguation. Our traversal algorithm leverages the context using co-occurrence and joint-matching, but convergence is not always guaranteed. Since the augmentation must also take place during inference time, we need convergence to be quick.

Lastly, we noticed that ConceptNet often places the same weight for seemingly unequally commonsensical relations between nodes in the graph. As with most crowd-sourced, some labeled information in ConceptNet is not reliable. However, because of its dynamic nature, its rapid growth, and popularity, we believe that these relations in the graph will contain more accurate weights and knowledge with time. Therefore, our work on incorporating ConceptNet lays out the ground work for a dynamic augmentation scheme.

6 Related Work

Most of the approaches today that try to embed commonsense reasoning or external knowledge for open-domain dialogue systems rely on a sequence to sequence framework that uses a LSTM or a GRU to encode dialogue context and the external information to be infused in the generated response (Parthasarathi and Pineau, 2018; Zhou et al., 2018; Zhang et al., 2020a; Young et al., 2017; Ghazvininejad et al., 2017; Moon et al., 2019; Wang et al., 2020). These approaches mainly differ in how they encode the commonsense KGs and combine them with the encoded dialogue text, such as using multiple LSTMs for a retrieval-based approach (Young et al., 2017) or using a graph attention mechanism to augment hidden representations of each word in the previous dialogue turn (Zhou et al., 2018). Others introduce additional related training tasks to become more sensitive to the factual source, such as auto-encoding tasks for the factual statements (Ghazvininejad et al., 2017) or using knowledge base question and answering tasks to enhance dialogue understanding and knowledge selection. Except for some of these works that evalute their work on multi-turn dialogue (Moon et al., 2019), most are limited in their assessment to single-turn responses (Young et al., 2017; Wang et al., 2020; Zhou et al., 2018; Ghazvininejad et al., 2017; Wu et al., 2020; Zhang et al., 2020a).

Wu et al. (2020) adapted the GPT-2 model (Radford et al., 2019) with sparse attention connections based on lexical control phrases and groundings for controllable and grounded response generation. However, the groundings are provided by Wikipedia pages that are linked to the conversation taking place, and therefore does not take advantage of existing knowledge graphs.

7 Conclusion

We present **AALK**, a modular dialogue data augmentation approach for developing a commonsense-aware and turn-level adaptive conversational dialogue dataset. Our preliminary experiments with the **AALK**-augmented Persona Chat dataset and fine-tuned *TransferTransfo* models show inconclusive results, but we identify limitations and sources of ineffectiveness that we will tackle in order to improve our augmentation framework.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 690–695, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020. Improving knowledge-aware dialogue generation via knowledge base question answering. In *AAAI*, pages 9169–9176.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, et al. 2020. A controllable model of grounded response generation. *arXiv preprint arXiv:2005.00613*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Tom Young, Erik Cambria, Iti Chaturvedi, Minlie Huang, Hao Zhou, and Subham Biswas. 2017. Augmenting end-to-end dialog systems with commonsense knowledge. *arXiv preprint arXiv:1709.05453*.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020a. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.