Saying No is An Art: Contextualized Fallback Responses for Unanswerable Dialogue Queries

Ashish Shrivastava[1], Kaustubh D. Dhole[1], Abhinav Bhatt[2], Sharvani Raghunath[1]
[1]Amelia Science, IPsoft R&D   [2]Universität des Saarlandes
[1]{firstname.lastname}@ipsoft.com
[2]abbh00001@stud.uni-saarland.de

Abstract

Despite end-to-end neural systems making significant progress in the last decade for task-oriented as well as chit-chat based dialogue systems, most dialogue systems rely on hybrid approaches which use a combination of rule-based, retrieval and generative approaches for generating a set of ranked responses. Such dialogue systems need to rely on a fallback mechanism to respond to out-of-domain or novel user queries which are not answerable within the scope of the dialog system. While, dialog systems today rely on static and unnatural responses like "I don't know the answer to that question" or "I'm not sure about that", we design a neural approach which generates responses which are contextually aware with the user query as well as say no to the user. Such customized responses provide paraphrasing ability and contextualization as well as improve the interaction with the user and reduce dialogue monotonicity. Our simple approach makes use of rules over dependency parses and a text-to-text transformer fine-tuned on synthetic data of question-response pairs generating highly relevant, grammatical as well as diverse questions. We perform automatic and manual evaluations to demonstrate the efficacy of the system.

1 Introduction

In order to cater to the diversity of questions spanning across various domains, dialogue systems generally follow a hybrid architecture wherein an ensemble of individual response subsystems (Kuratov et al.; Harrison et al., 2020) are employed from which an appropriate response is presented to the user (Serban et al., 2017; Finch et al., 2020; Paranjape et al., 2020). However, it is common for dialogue systems to encounter queries which are not within their scope of knowledge. While increasing the number of such subsystems would be

Figure 1: Comparison of responses of three flight booking dialog systems: The first one does not handle unknown responses. The second one has a default fallback response. The third one has a fall-back response which is contextualized with the user query.

a good strategy to increase coverage, it can be a never ending process and a default fallback strategy would always be needed. Besides, domain specific dialog systems, especially those deployed in professional settings generally prefer restricting themselves to a fixed set of domains, and purposely refrain from responding to out-of-domain and random or toxic user queries.

One approach to acknowledge such queries is to have a fallback mechanism with responses like "I don't know the answer to this question" or "I'm not sure how to answer that." However, such responses are static and unengaging and give an impression that the user's query has gone unacknowledged or is not understood by the system as shown in Figure 1 above.

Yu et al. (2016) have shown that static and predefined responses lead to lower levels of user engagement and decrease users' interest in interacting with the system.

Our fallback approach attempts to address these limitations by generating "don't-know" responses which are engaging and contextually closer with the user query. 1) Since there are no publicly available datasets to generate such contextualised responses, we synthetically generate (query, fall-

back response) pairs using a set of highly accurate handcrafted dependency patterns. 2) We then train a sequence-to-sequence model over synthetic and natural paraphrases of these queries. 3) Finally, we measure the grammaticality and relevance of our models using a crowd-sourced setting to assess the generation capability. We publicly release the code and training data used in our experiments. [1]

2 Related Work

Improving the coverage to address out-of-domain queries is not a new problem in designing dialog systems. The most popular approach has been via presenting the user with chit-chat responses. Other systems such as Blender (Roller et al., 2020) and Meena (Adiwardana et al., 2020) promise to be successful for open-domain settings. Paranjape et al. (2020) finetune a GPT-2 model (Radford et al., 2019) on the EmpatheticDialogues dataset (Rashkin et al., 2019) to generate social talk responses. While this might seem fitting for chit-chat and social talk dialog systems, domain-specific scenarios often dealing with professional settings would refrain from performing friendly or social talk especially avoiding the possibility of the randomness of generative models. Also, multiple subsystem architectures always have the possibility of cascading errors and profane or toxic queries. Hence systems always have a foolproof mechanism in the form of static templates to reply from. Liang et al. (2020) uses an interesting approach for error handling by mapping dialog acts and intents to templates. Besides, like Finch et al. (2020) it is always safer to generate fallback responses on encountering queries which might be toxic, biased or profane. [2]

Another line of work attempts to handle user queries which are ambiguous by asking back clarification questions (Dhole, 2020; Zamani et al., 2020; Yu et al., 2020). While this increases user interaction and coverage to an appreciable extent, it does not eliminate the requirement of a failsafe fallback responder. This paper's contribution is to address this requirement with an enhanced version of a fallback response generator.

3 Methods

We describe two approaches to generate such contextual don't-know responses.

3.1 The Dependency Based Approach (DBA)

Inspired by previous approaches which use parse structures to generate questions (Heilman and Smith, 2009; Mazidi and Tarau, 2016; Dhole and Manning, 2020), we create a rule-based generator by handcrafting dependency templates to cater to a wide variety of question patterns as shown in Figure 2. We perform extensive manual testing to improve the generations from these rules and increase overall coverage. The purpose of these rules is two-fold: i) To create a high-precision fall-back response generator as a baseline and ii) to help create (query, don't-know-response) pairs which could be paired with natural paraphrases to serve as seed training data for other deep learning architectures.

To build this baseline generator, we utilize 9 dependency templates in the style of SynQG (Dhole and Manning, 2020). We utilize the dependency parser from Andor et al. (2016) to get the Universal Dependencies (Nivre et al., 2016, 2017) of the user query. We then convert it to a don't-know-response by re-arranging nodes to a matched template. We further change pronouns, incorporate named entity information, and add rules to handle modals and auxiliaries. Finally, we also add rules for flipping pronouns to convert an agent targeted question to a user targeted response by interchanging pronouns and their supporting verbs. E.g. You to I and vice-versa.

We incorporate a bit of paraphrasing by randomizing various prefixes like "I'm not sure whether", "I don't know if", etc. and randomly using named entities. We describe the high-level algorithm below and in Algorithm 1.

$$prefix = pickRandom(prefixPool)$$
$$response = DBR(Question)$$
$$suffix = pickRandom(suffixPool)$$
$$fallbackResponse = Concat(prefix,$$
$$response, suffix)$$

3.2 Sequence-to-Sequence Approach

Owing to the expected low coverage and scalability of the rule-based approach, we resort to take

---

[2] Handling programming exceptions and code failures also necessitates a simple fallback approach.

Figure 2: Dependency Rules with the class of questions they cater too and their corresponding responses. In the second last sentence, the named entity "Will" is randomly replaced by "that person".

Algorithm 1 Dependency Based Response (DBR)

$nodes \leftarrow dependencyParse(Question)$
for each template in templatePool **do**
    if (template condition matched) **then**
        Populate template using *nodes*
        Handle modals & auxilliaries
        Flip pronoun
        Randomly substitute Named Entity
if no template condition matched **then**
    $return\ pickRandom(\text{defaultResponse pool})$
$return$ filled template response

---

advantage of pre-trained neural architectures to attempt to create a sequence-to-sequence fallback responder. To incorporate noise and avoid the model to over-fit on the handcrafted transformations, we do not train the model directly on (query, don't-know-response) pairs generated from the previous section. From all possible questions of the Quora Questions Pairs dataset (QQP) [3], we first filter all the questions which generate a reply from the dependency based rules. Then we pair these dont-know-responses with the paraphrases of the input questions rather than the input questions themselves. [4] Primarily attempting to avoid over-fitting on the dependency patterns, this also helps generate dont-know-responses which are paraphrastic in nature.

Metrics	DBA	Seq-To-Seq
%GC	81.6	87.2
ARS	3.97	3.66

Table 1: Human evaluation between the two approaches. %GC= % of Grammatically correct responses, ARS=Average Relevance Score.

After incorporating paraphrases from QQP, we

are able to build a dataset of 100k pairs, which we call the "I Dont Know Dataset" (IDKD). After witnessing the success of text-to-text transformers, we use the pre-trained T5 transformer (Raffel et al., 2020) as our sequence-to-sequence model. We divide IDKD into a train and validation split of 80:20. We use the Transformers code from HuggingFace (Wolf et al., 2020) to fine-tune a T5-base model over IDKD for 2 epochs. [5]

4 Results

Most prior generated systems are evaluated on a range of automatic metrics like BLEU and ROGUE (Papineni et al., 2002) used in the machine translation literature. However, owing to the drawbacks of these metrics, we perform human evaluation of the generated responses using two metrics - namely "relevance" and "grammaticality" as defined in Dhole and Manning (2020). We evaluate the performance of both the approaches in a crowd-sourced setting by requesting English-schooled individuals to rate. [6] Raters were asked to evaluate grammaticality in a binary setting (grammatical/ungrammatical) and relevance on a Likert scale (1 to 5).

Our human evaluations are shown in Table-1. T5 responses tend to be more grammatical than their dependency counterparts by a large margin of 6%. Relevance scores drop slightly from 3.97 to 3.66. This can be largely attributed to the model's paraphrastic ability of describing words and connected events outside the knowledge of the user's query. Eg. in the second query in Table 3, if the string "MIT" were something other than an institution, the dependency based approach would seem

---

[3]Quora Question Pairs Dataset
[4]Those question pairs which have the label "1" or are similar are used as paraphrases.

[5]Increasing the epochs to 3 and 4 tended to overfit and not generate paraphrases.
[6]Our pool of English-schooled raters consisted of engineering graduates having attained high school and engineering education with English as a medium of instruction.

Metrics	Question	Dependency Based Approach	Seq2Seq Approach
Coverage	–	54.56%	100%
Average Sentence Length	9.41	11.83	11.78
Sentence Length Variance	11.74	11.60	10.66
Average #Novel Words	–	0.91	2.57

Table 2: System evaluation between Dependency Based Approach and Sequence to Sequence Approach.

Question: What is the quickest way to increase Instagram followers?
DBA: I am not sure what the quickest way to increase Instagram followers is.
Seq2Seq: I cannot be sure how a user can increase Instagram followers.

Question: How can I join MIT?
DBA: I am not sure how you can join MIT.
Seq2Seq: I am not sure what the eligibility criteria for MIT are.

Question: How can I drive more traffic to my website?
DBA: I'm not really sure how you can drive more traffic.
Seq2Seq: I am not sure what the best way to get free traffic to your website is.

Question: Where can I find a collection of poetry by William Shakespeare?
DBA: I'm not really sure where you can find a collection of poetry by William Shakespeare.
Seq2Seq: I'm not actually sure where you can get a collection of his poems.

Question: What is the effect of demonetization of 500 and 1000 rupees notes?
DBA: I cannot be sure what the effect of demonetization of 500 and 1000 rupees notes is.
Seq2Seq: I'm not really sure what my views on PM Modi's decision on discontinuing 500 & 1000 rs notes are.

Table 3: Natural don't know responses generated by both the approaches. Highlights in blue depict words, phrases or events not mentioned by the user.

safer than the seq2seq approach.

In addition, T5 responses on an average generate at least double the number of novel words than their dependency counterparts as shown in Table 2. Sentence length mostly remains unaffected across the two models. Undoubtedly, the rule-based model despite being highly relevant is only able to reply to 54.5% of random QQP queries.

The T5 model helped to not only add paraphrastic variations but also scale to user queries outside of the scope of the dependency templates. More importantly, without losing the original ability of saying no, the model was able to generate more natural sounding dont-know-reponses by utilizing it's inherent world-knowledge acquired during pre-training. Table 3 shows some interesting examples. The highlighted phrases in blue show the benefits of the model's pre-training ability.

5   Conclusion and Future work

We describe two simple approaches which enhance user interaction to cater to the necessities of real-life dialogue systems which are generally a tapestry of multiple solitary subsystems. In order to avoid cascading errors from such systems, as well as refrain from answering out-of-domain and toxic queries it is but natural to have a fallback approach to say no. We argue that such a fallback approach could be contextualised to generate engaging responses by having multiple ways of saying no rather than a one common string for all approach. The appeal of our approach is the ease with which it can rightly fit within any larger dialog design framework.

Of course, this is not to deny that as we give more paraphrasing power to the fallback system, it would tend to retract from succinctly replying with a no - as is evident from the drop in the relevance scores. Nevertheless, we still believe that both our fallback approaches could serve as effective baselines for future work.

6 Acknowledgments

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*.

Kaustubh Dhole and Christopher D. Manning. 2020. Syn-QG: Syntactic and shallow semantic rules for question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 752–765, Online. Association for Computational Linguistics.

Kaustubh D Dhole. 2020. Resolving intent ambiguities by retrieving discriminative clarifying questions. *arXiv preprint arXiv:2008.07559*.

Sarah E Finch, James D Finch, Ali Ahmadvand, Xiangjue Dong, Ruixiang Qi, Harshita Sahijwani, Sergey Volokhin, Zihan Wang, Zihao Wang, Jinho D Choi, et al. 2020. Emora: An inquisitive social chatbot who cares for you. *arXiv preprint arXiv:2009.04617*.

Vrindavan Harrison, Juraj Juraska, Wen Cui, Lena Reed, Kevin K Bowden, Jiaqi Wu, Brian Schwarzmann, Abteen Ebrahimi, Rishi Rajasekaran, Nikhil Varghese, et al. 2020. Athena: Constructing dialogues dynamically with discourse constraints. *arXiv preprint arXiv:2011.10683*.

Michael Heilman and Noah A Smith. 2009. Question generation via overgenerating transformations and ranking. Technical report, Carnegie-Mellon Univ Pittsburgh pa language technologies insT.

Yuri Kuratov, Idris Yusupov, Dilyara Baymurzina, Denis Kuznetsov, Daniil Cherniavskii, Alexander Dmitrievskiy, Elena Ermakova, Fedor Ignatov, Dmitry Karpov, and Daniel Kornev. Dream technical report for the alexa prize 2019.

Kaihui Liang, Austin Chau, Yu Li, Xueyuan Lu, Dian Yu, Mingyang Zhou, Ishan Jain, Sam Davidson, Josh Arnold, Minh Nguyen, et al. 2020. Gunrock 2.0: A user adaptive social conversational system. *arXiv preprint arXiv:2011.08906*.

Karen Mazidi and Paul Tarau. 2016. Infusing NLU into automatic question generation. In *Proceedings of the 9th International Natural Language Generation conference*, pages 51–60, Edinburgh, UK. Association for Computational Linguistics.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, et al. 2017. Universal dependencies 2.1.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

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.

Ashwin Paranjape, Abigail See, Kathleen Kenealy, Haojun Li, Amelia Hardy, Peng Qi, Kaushik Ram Sadagopan, Nguyet Minh Phu, Dilara Soylu, and Christopher D Manning. 2020. Neural generation meets real people: Towards emotionally engaging mixed-initiative conversations. *arXiv preprint arXiv:2008.12348*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

Iulian V Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. 2017. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lili Yu, Howard Chen, Sida I. Wang, Tao Lei, and Yoav Artzi. 2020. Interactive classification by asking informative questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2664–2680, Online. Association for Computational Linguistics.

Zhou Yu, Leah Nicolich-Henkin, Alan W Black, and Alexander Rudnicky. 2016. A Wizard-of-Oz study on a non-task-oriented dialog systems that reacts to user engagement. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 55–63, Los Angeles. Association for Computational Linguistics.

Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, pages 418–428.