An Empirical study to understand the Compositional Prowess of Neural Dialog Models

Vinayshekhar Bannihatti Kumar*
Applied Scientist AWS AI
vinayshk@amazon.com

Vaibhav Kumar*
Applied Scientist Alexa AI
kvabh@amazon.com

Mukul Bhutani
Carnegie Mellon University
mukul.bhutani93@gmail.com

Alexander Rudnicky
Carnegie Mellon University
air@cmu.edu

Abstract

In this work, we examine the problems associated with neural dialog models under the common theme of compositionality. Specifically, we investigate three manifestations of compositionality: (1) Productivity, (2) Substitutivity, and (3) Systematicity. These manifestations shed light on the generalization, syntactic robustness, and semantic capabilities of neural dialog models. We design probing experiments by perturbing the training data to study the above phenomenon. We make informative observations based on automated metrics and hope that this work increases research interest in understanding the capacity of these models.

1 Introduction

Fully data-driven and end-to-end approaches to dialog response generation (Vinyals and Le, 2015; Serban et al., 2016; Bordes et al., 2016; Serban et al., 2017; Zhao et al., 2017) within the sequence-to-sequence (seq2seq) (Hochreiter and Schmidhuber, 1997; Sutskever et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017) framework have become ubiquitous and now produce competitive results.

Recently, there have been a few attempts to explore the capabilities of such models. A well known problem in seq2seq modeling is the tendency to generate short and meaningless replies in conversation (Li et al., 2015; Mou et al., 2016). By drawing a parallel between machine translation and dialog generation, Wei et al. (2019) suggest that such models encounter a severe mis-alignment problem i.e a given input utterance can have many plausible replies.

Sankar et al. (2019) empirically investigate the information captured in seq2seq models by synthetically perturbing the test set during inference. They demonstrate an inability of seq2seq models to use all the information that is presented. They also present their study as a "diagnostic tool" to evaluate dialog models.

Although they provide useful insights, such studies fail to systematically demonstrate the compositional features of seq2seq dialog models. Further, their "diagnostic tool" is only helpful for evaluating syntactic robustness of models at *test time*. In this work, we carefully design experiments to investigate and evaluate the compositional generalizability of neural dialog models.

Compositionality has been well studied for Neural Machine Translation (Cho et al., 2014; Lake and Baroni, 2017) as well as some other tasks. In these works, for a system to be compositional, it should be able to generalize beyond its observations. For example, Kaiser and Sutskever (2015) observe that Neural GPUs are able to generalize addition and multiplication to larger sequences than what they are trained on. However, one should carefully note that such a definition of compositionality is peripheral and represents only a part of what it truly means.

To provide a complete picture, Hupkes et al. (2019) collect the different manifestations of compositionality and translate them into a series of theoretically-grounded tests. By adapting (and modifying) some of these tests, the experiments in this paper aim to quantitatively elucidate the compositional nature of seq2seq based neural dialog models. Below, we provide a motivation and description for each of the adapted tests:

Productivity - Upon taking part in a number of reasonable length conversations, it might not be difficult for humans to carry conversations consisting of a larger number of turns. Based on this intuition, we test the ability of a dialog system to **extend its prediction** beyond the length of the observed conversational history.

Substitutivity - There is a many-to-many correspondence between utterances and their possible responses. Given the responses of a particular con-

* equal contribution, Work done while students at CMU

Proceedings of the Third Workshop on Insights from Negative Results in NLP, pages 154 - 158

Dataset	Baseline	DS-0.75	DS-0.5	DS-0.25	DNS-0.75	DNS-0.5	DNS-0.25	BT-Russian
				Transformer				
dailydialog	$33.2_{[0.7]}$	$140.6_{[11.6]}$	$56.3_{[2.0]}$	$41.1_{[1.3]}$	$131.6_{[2.6]}$	$63.4_{[1.0]}$	$42.9_{[0.4]}$	$117.2_{[7.9]}$
MutualFriends	$12.5_{[0.1]}$	$30.1_{[3.0]}$	$18.1_{[1.2]}$	$15.0_{[0.3]}$	$39.6_{[1.1]}$	$21.3_{[0.8]}$	$17.1_{[0.3]}$	$150.8_{[16.8]}$
Babi	$1.0_{[0.0]}$	$19.8_{[0.7]}$	$6.3_{[0.4]}$	$3.5_{[0.2]}$	$16.1_{[1.6]}$	$3.3_{[0.1]}$	$2.1_{[0.1]}$	$6.4_{[1.2]}$
				S2S				
dailydialog	$29.4_{[0.3]}$	$104.8_{[2.4]}$	$47.1_{[0.6]}$	$35.6_{[0.2]}$	$150.9_{[5.4]}$	$61.9_{[1.3]}$	$39.4_{[0.5]}$	$192.9_{[14.3]}$
MutualFriends	$13.3_{[0.1]}$	$25.4_{[0.2]}$	$17.2_{[0.2]}$	$15.2_{[0.3]}$	$50.1_{[2.1]}$	$24.3_{[0.5]}$	$18.3_{[0.3]}$	$227.1_{[8.6]}$
Babi	$1.2_{[0.0]}$	$3759.0_{[1994.7]}$	$52.6_{[13.2]}$	$8.2_{[1.4]}$	$121.0_{[24.4]}$	$7.9_{[1.8]}$	$3.0_{[0.1]}$	$59.3_{[14.9]}$
				S2SA				
dailydialog	$26.9_{[0.2]}$	$94.7_{[4.0]}$	$45.5_{[0.2]}$	$32.6_{[0.3]}$	$130.2_{[5.3]}$	$58.6_{[1.1]}$	$37.3_{[0.7]}$	$173.0_{[16.5]}$
MutualFriends	$10.2_{[0.1]}$	$20.1_{[0.3]}$	$13.6_{[0.1]}$	$11.8_{[0.2]}$	$40.5_{[1.4]}$	$19.0_{[0.2]}$	$14.1_{[0.2]}$	$216.4_{[18.4]}$
Babi	$1.0_{[0.0]}$	$961.0_{[421.5]}$	$68.2_{[22.5]}$	$8.1_{[2.2]}$	$118.8_{[43.4]}$	$7.5_{[1.2]}$	$2.8_{[0.2]}$	$630.8_{[136.1]}$

Table 1: Performance of the models based on perplexity. The second column represents the baseline scores of the models on different datasets. Columns 3-5 shows the effect of dropping stop words at a certain rate. Columns 6-8 shows the effect of dropping non stop words at a certain rate. Column 9 shows the difference in perplexity of the model when the test set is changed by back translation and evaluated using the baseline model. All experiments are repeated 5 times and the mean(μ) and std deviations(σ) are reported in every cell. For all experiment runs and other metrics refer to A.1.

versation, if we encounter a semantically equivalent conversation, we can easily produce the same set of responses to this new conversation. Based on this, we attempt to observe if dialog models are also capable of such reasoning. This property of compositionality accounts for the **semantic expressiveness** of neural models.

Systematicity - Humans can understand how to fill in missing pieces of information, or to introduce additional words which can make an utterance in a conversation more fluent. This makes humans capable of recombining known fragments and rules. Without the presence of topic-inducing words, it might become difficult for humans to make sense of a conversation. Based on this intuition, we test the ability of the model to recombine known fragments and rules. This property of compositionality accounts for the **syntactic robustness** of neural models.

The contributions of this paper are threefold: **(i)** We observe that neural dialog models don't generalize well to dialogs with longer turns when they are trained on dialogs with shorter number of turns. **(ii)** Neural dialog models pay less attention to the topic inducing "content words" of the dialog. In fact, we observe that they are highly sensitive to the stop words (a type of "function word") present in utterances. **(iii)** We also observe that the neural dialog models don't perform well when the same utterance is presented to the model in a semantically similar but syntactically different fashion i.e they are not robust to syntactic variations. The code for reproducing results is released along with this paper [1].

2 Datasets

Following Sankar et al. (2019), we experiment with using an open domain, a closed domain, and a synthetically generated dataset. The details of the dataset are presented below:

DailyDialog: An open domain, manually labelled dataset (Li et al., 2017) consisting of conversations on multiple topics which can occur on a daily basis. There are 13,118 total dialogs with an average of 7.9 turns per dialog.

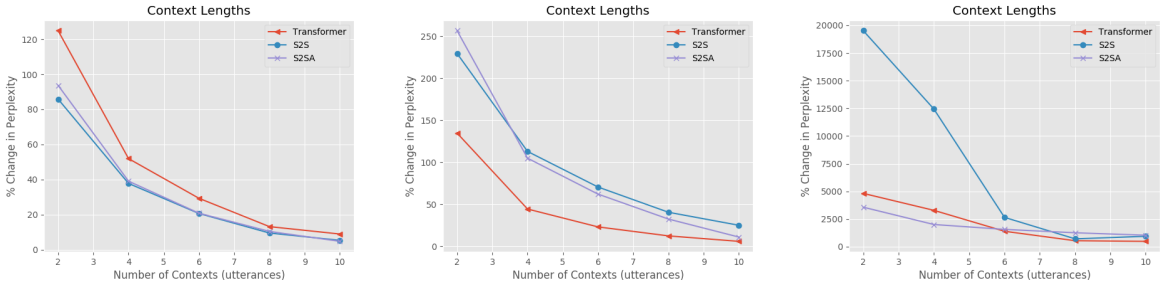
Mutual Friends: A task-oriented dataset (He et al., 2017) that encourages open-ended dialog acts. It has a total of 11,157 dialogs with an average length of 11.4 utterances per dialog.

Babi: A synthetic dataset created by Bordes et al. (2016). We use task 5 of this dataset which requires the prediction of the text of the entire dialog and not just dialog acts. Each dialog in this task has an average of 13 utterances and there is a total of 1,000 dialogs.

3 Experiments and Results

We investigate using Seq2Seq(**S2S**) (Sutskever et al., 2014), Seq2Seq-Attention(**S2SA**) (Luong et al., 2015) and Transformer models (Vaswani

[1] https://github.com/vinayshekharcmu/ComposionalityOfDialogModels



(a) Daily Dialog (b) Mutual Friends (c) Babi - Task 5

Figure 1: Results of the model on the test of productivity. We see that all the models don't learn to generalize from dialogs with fewer utterances to dialogs with more utterances.

et al., 2017). The behaviour of these models is examined using the three standard datasets described in Section 2.

Both S2S and S2SA utilise a two-layer LSTM for the encoder and the decoder. Each layer has 128 hidden units with a dropout of 0.1. On the other hand, the transformer utilises a 300 dimensional embedding with 2 layers and 2 attention heads. Perplexity has been shown to correlate well with **human judgement** for Dialog Systems (Adiwardana et al., 2020) making it a suitable metric for our study. By choosing perplexity we also remain consistent with the previous study conducted by Sankar et al. (2019). Note that we do not aim to achieve state-of-the-art results, but rather, our aim is to observe and characterize the behaviour of the models based on different aspects of compositionality. Hence we pick three seminal models that tackles the problem of language generation and probe them to understand their manifestations.

The upcoming subsections first provides a brief description of the experimental setup employed for measuring the compositional capabilities of the various models, and then later discusses the results.

3.1 Productivity

This experiment aims to test whether neural dialog models can learn from meaningful dialogs consisting of fewer utterances and then generalize to dialogs consisting of a larger number of utterances than what they had observed during training time.

In order to test this capability, we train the models with trimmed context. For each dialog in the training set, we restrict the context utilised by the models to the previous k utterances, where $k \in \{2, 4, 6, 8, 10\}$. However, at test time the models utilise all the available context. We compare the performance of the models trained on different context lengths to that of the baseline model which is trained by utilising the entire context.

The results are displayed in Figures 1a, 1b, 1c. These figures show the % increase in perplexity of the models from their baseline perplexity as a function of number of utterances in the dialog. It is quite clear from the figures that the model are incapable of generalizing from shorter dialogs to longer dialogs.

The average number of utterances within the dialogs is $\sim$ 8 for all the three datasets. Based on the results we see that even when models use previous 8 utterances, their performance is still significantly lower than that of the baseline. This experiment questions the generalizing ability of the model beyond what was observed during train time.

3.2 Systematicity

Two different experiments were performed to understand the semantic robustness of these models. The first experiment was done to understand the importance of stop words. A comparison between model's sensitivity to dropping of stop words (**DS**) and dropping of content words (**DNS**) sheds light on the relevance of stop words in dialogs. We drop stop words and content words at the rate of 0.75, 0.5 and 0.25 and observe the effect on models' performance. When the rate of stop words removal is 1, all the stop words are removed and when it is 0.25, 25% are removed, etc.

In second experiment we drop words based on their rank in the corpus. Six different conditions are used in this experiment. We first drop words from the top ranks such that only 10% of the total number of words are removed in the corpus. We then repeat this by using the mid ranked words. Ideally, the models should be affected equally in

Rank Range	Transformer	S2S	S2SA
	DailyDialog		
0-1	$49.4_{[0.6]}$	$49.1_{[7.1]}$	$42.4_{[1.3]}$
1-3	$59.6_{[0.9]}$	$59.5_{[2.1]}$	$55.4_{[1.6]}$
0-3	$92.7_{[1.9]}$	$92.5_{[1.3]}$	$88.7_{[4.2]}$
500-1000	$52.6_{[1.1]}$	$59.6_{[1.1]}$	$52.3_{[1.2]}$
1000-1500	$39.4_{[0.7]}$	$42.0_{[1.1]}$	$38.8_{[0.4]}$
500-1500	$59.5_{[0.9]}$	$76.3_{[3.6]}$	$72.5_{[1.8]}$
	MutualFriends		
0-1	$14.9_{[0.2]}$	$16.7_{[0.5]}$	$12.9_{[0.5]}$
1-3	$17.6_{[0.4]}$	$20.5_{[0.4]}$	$15.0_{[0.3]}$
0-3	$19.9_{[0.4]}$	$23.0_{[0.5]}$	$18.0_{[0.8]}$
300-600	$13.9_{[0.2]}$	$15.1_{[0.2]}$	$12.0_{[0.8]}$
600-1000	$14.3_{[0.3]}$	$15.1_{[0.2]}$	$11.8_{[0.3]}$
300-1000	$16.0_{[0.5]}$	$17.8_{[0.2]}$	$13.8_{[0.2]}$
	Babi		
0-1	$2.0_{[0.1]}$	$3.9_{[0.5]}$	$4.8_{[0.7]}$
0-2	$3.7_{[0.2]}$	$11.2_{[1.4]}$	$11.0_{[1.6]}$
36-44	$1.5_{[0.0]}$	$2.1_{[0.1]}$	$2.0_{[0.1]}$
36-55	$1.5_{[0.0]}$	$2.2_{[0.1]}$	$2.1_{[0.1]}$

Table 2: The first column represents the range of ranks based on which the words were removed from the dataset. We chose to experiment with the top and the mid ranking words. We dropped words from both sections such that it accounts for $\approx 10\%$ of the words in its respective corpus. We see that the model is very sensitive to the top ranked words (which are stop-words most of the time). The effect of dropping 1000, 700 and 20 "content words" from the middle section is equivalent to dropping 3,3,2 stop words for dailydialog, mutualfriends and babi respectively.

both these settings, as, in each setting we end up removing 10% of the words in the training data. In fact, it should be affected more in the latter case as the mid-rank words are majorly responsible for inducing the topic of the dialog and it should be difficult to continue a conversation without knowing the topic. Note that, for both these experiments, we do not remove any word during test time.

Table 1 shows the result of the first experiment. We see that each of model's performance increases as the rate of dropping stop words decreases. This observation suggests the high sensitivty of the models towards stop words. Even dropping 25% of the stop words affects the models adversely. While dropping of the content words also affects models performance, we observe that all the models perform just slightly worse when content words are dropped as compared to stop words. However, it is interesting to see that the transformer's performance is stable across different drop rates whereas the LSTM based sequence to sequence models suffer when the drop rate is high.

The results for the second experiment are provided in Table 2. It is clear that removal of higher ranked words leads to a greater drop in the model performance when compared to the drop caused by the removal of middle ranked words, even though in both the cases we remove the same percentage of words. This provides two insights: (1) Models don't focus on the mid ranking words (which are mostly topic inducing) and (2) Models have an over-reliance on top ranking words (which are mostly stop words).

3.3 Substitutivity

Given that we (humans) know the answer to a particular question, we will not have any difficulty in answering it even if it is asked in various different ways. This experiment aims to test if neural dialog models are also capable of this ability.

In order to do this, we evaluate the baseline models on the backtranslated (**BT**) version of the test set. Basically, back translation provides a paraphrased version of individual utterances (Wieting et al., 2017), which brings in syntactic variations while keeping the semantics intact.

We back translate the test set from both German and Russian back into English. Since the BLEU scores when translating from German were considerably lower than that of Russian, we decided to test the models based on Russian Backtranslations. The final backtranslations have a BLEU score of 35.91, 10.12, 43.49 on Daily Dialog, Mutual Friends and Babi respectively.

The results for the experiment are provided in Table 1. It is clear that the models are adversely affected when presented with back translated (paraphrased) utterances. One would expect the models to have similar perplexities when utterances are paraphrased, however we see that there is a significant increase in perplexity. This observation is consistent across the three different models. We also observe that the transformer is slightly more robust to syntactic variation than others.

4 Conclusion

This work interprets the behaviour of seq2seq based Neural Dialog Models under the general umbrella of compositionality. We observe that such models lack the ability to reason and produce response based on surface level information. The results

provided in this paper motivate the need for better modelling approaches.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. *arXiv preprint arXiv:1704.07130*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2019. The compositionality of neural networks: integrating symbolism and connectionism. *arXiv preprint arXiv:1908.08351*.

Łukasz Kaiser and Ilya Sutskever. 2015. Neural gpus learn algorithms. *arXiv preprint arXiv:1511.08228*.

Brenden M Lake and Marco Baroni. 2017. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *arXiv preprint arXiv:1711.00350*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. *arXiv preprint arXiv:1607.00970*.

Chinnadhurai Sankar, Sandeep Subramanian, Christopher Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. *arXiv preprint arXiv:1906.01603*.

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*.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

I Sutskever, O Vinyals, and QV Le. 2014. Sequence to sequence learning with neural networks. *Advances in NIPS*.

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

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Bolin Wei, Shuai Lu, Lili Mou, Hao Zhou, Pascal Poupart, Ge Li, and Zhi Jin. 2019. Why do neural dialog systems generate short and meaningless replies? a comparison between dialog and translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7290–7294. IEEE.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. *arXiv preprint arXiv:1706.01847*.

Tiancheng Zhao, Allen Lu, Kyusong Lee, and Maxine Eskenazi. 2017. Generative encoder-decoder models for task-oriented spoken dialog systems with chatting capability. *arXiv preprint arXiv:1706.08476*.