WHY DO NEURAL DIALOG SYSTEMS GENERATE SHORT AND MEANINGLESS REPLIES? A COMPARISON BETWEEN DIALOG AND TRANSLATION

Bolin Wei,[†] *Shuai Lu,*[†] *Lili Mou,*[‡] *Hao Zhou,*[¶] *Pascal Poupart,*[‡] *Ge Li,*[†] *Zhi Jin*[†]

[†]Key Laboratory of High Confidence Software Technologies (Peking University)
Ministry of Education, China; Software Institute, Peking University, China
[‡]University of Waterloo [¶]Toutiao AI Lab

ABSTRACT

This paper addresses the question: In neural dialog systems, why do sequence-to-sequence (Seq2Seq) neural networks generate short and meaningless replies for open-domain response generation? We conjecture that in a dialog system, due to the randomness of spoken language, there may be multiple equally plausible replies for one utterance, causing the deficiency of a Seq2Seq model. To evaluate our conjecture, we propose a systematic way to mimic the dialog scenario in machine translation systems with both real datasets and toy datasets generated elaborately. Experimental results show that we manage to reproduce the phenomenon of generating short and meaningless sentences in the translation setting.

Index Terms— Dialog system, machine translation, sequence-to-sequence, short replies

1. INTRODUCTION

Open-domain human-computer dialog systems are attracting increasing attention in the NLP community. With the development of deep learning, sequence-to-sequence (Seq2Seq) neural networks or more generally encoder-decoder frameworks, are among the most popular models for text-based response generation in dialog systems [1, 2, 3, 4].

Historically, Seq2Seq-like models are first designed for machine translation [5, 6] and later widely applied to image captioning [7], text summarization [8], etc. When adapted to text-based open-domain dialog systems, however, Seq2Seq models are less satisfactory. A severe problem is that the Seq2Seq model tends to generate short and meaningless replies, e.g., "I don't know" [2] and "Me too" [3]. They are universally relevant to most utterances, called *universal replies* in [3], and hence less desired in real-world conversation systems.

In previous studies, researchers have proposed a variety of approaches to address the problem of universal replies, ranging from heuristically modified training objectives [2], diversified decoding algorithms [9], to content-introducing approaches [3, 10]. Although the problems of universal replies have been alleviated to some extent, there lacks an empirical explanation to the curious question: *Why does the same Seq2Seq model tend to generate shorter and less meaningful sentences in a dialog system than in a machine translation system?*

Considering the difference between dialog and translation data, our intuition is that, compared with translation data, a dialog system encounters a severe "unaligned" problem due to the randomness and uncertainty of spoken language: an utterance can be matched to multiple equally plausible replies, but these replies may have different meanings. On the contrary, the translation datasets typically have a more precise semantic matching between the source and target sides. This conjecture is casually expressed in previous work [3], but is so far not supported by experiments.

To verify our conjecture, we propose a method by mimicking the unaligned phenomenon on machine translation datasets, which is to shuffle the source and target sides of the translation pairs to artificially build a conditional distribution of target sentences with multiple plausible data points. We conduct experiments on a widely used translation dataset; we further conduct a simulation with some predefined distributions, serving as additional evidence. The experimental results show that shuffling of datasets tends to make translated sentences shorter and less meaningful. Therefore, the unaligned problem can be one reason that causes short and meaningless replies in neural dialog systems.

To summarize, this paper compares Seq2Seq dialog with translation systems, and provides an explanation to the question: Why do neural dialog systems tend to generate short and meaningless replies? Our study also sheds light on the future development of neural dialog systems as well as the application scenarios where Seq2Seq models are appropriate.

2. CONJECTURE

We hypothesize that *given a source sequence, the conditional distribution of the target sequence having multiple plausible points is one cause of the deficiency of Seq2Seq models in dialog systems*.

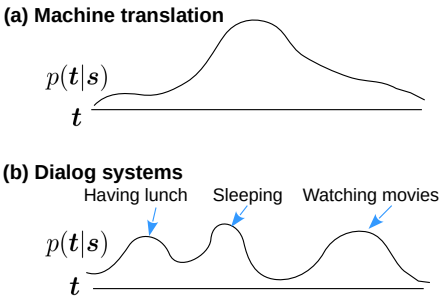

Fig. 1: The conditional distribution $p(\boldsymbol{t}|\boldsymbol{s})$ in (a) machine translation and (b) dialog systems, where we consider an analog of continuous random variables. More rigorously speaking, $p(\boldsymbol{t}|\boldsymbol{s})$ is peaked at one or a few similar sentence(s) in machine translation because source and target information generally aligns, whereas an utterance can have multiple plausible replies in dialog systems.

Let us denote the source sequence by $\boldsymbol{s} = s_1, s_2, \cdots, s_{|\boldsymbol{s}|}$ and the target sequence by $\boldsymbol{t} = t_1, t_2, \cdots, t_{|\boldsymbol{t}|}$. Both (orthodox) training and prediction objectives are to maximize $p_{\boldsymbol{\theta}}(\boldsymbol{t}|\boldsymbol{s})$, where the conditional probability $p_{\boldsymbol{\theta}}(\cdot|\cdot)$ is modeled by a Seq2Seq neural network with parameters $\boldsymbol{\theta}$.

In a machine translation system, the source and target information generally aligns well, although some meanings could have different expressions. Figure 1a shows a continuous analog of $p(\boldsymbol{t}|\boldsymbol{s})$.

In an open-domain dialog system, however, an utterance may have a variety of replies that are (nearly) equally plausible. For example, given a user-issued utterance "What are you going to do?" there could be multiple replies like "having lunch," "watching movies," and "sleeping," shown in Figure 1b with an analog using continuous variables. There is no particular reason why one reply should be favored over another without further context; even with context, this problem could not be fully solved because of the true randomness of dialog. Located near the "mode" could be viewed as replies of similar meanings but less fluent expressions. Other areas with low probabilities are nonsensical utterances that are either not fluent in spoken language or irrelevant to the previous utterance $\boldsymbol{s}$.

The above is, perhaps, the most salient difference between dialog and translation datasets. Although it is tempting to think of Seq2Seq's performance in this way [3], barely a practical approach exists to verify the conjecture in the dialog setting alone. In the rest of this paper, we will verify it with a modification of machine translation tasks.

3. EXPERIMENTAL PROTOCOL

3.1. Mimicking a "Dialog Scenario" in Translation

We propose to mimic the "unaligned" property in machine translation datasets by shuffling the source and target pairs. This ensures the resulting conditional distribution $p(\boldsymbol{t}|\boldsymbol{s})$ to have multiple plausible target sequences, whereas other settings of translation remain unchanged, making a rigorous controlled experiment.

Formally speaking, let $\{(\boldsymbol{s}^{(n)}, \boldsymbol{t}^{(n)})\}_{n=1}^N$ be the training dataset in a translation setting, where $(\boldsymbol{s}^{(n)}, \boldsymbol{t}^{(n)})$ is a particular data point containing a source and target sentence pair; in total we have N data points.

The shuffled dataset is $\{(\boldsymbol{s}^{(n)}, \widetilde{\boldsymbol{t}}^{(n)})\}_{n=1}^N$, where $\widetilde{\boldsymbol{t}}^{(n)} = \boldsymbol{t}^{(\tau(n))}$ and $\tau(1), \cdots, \tau(N)$ is a random permutation of $1, 2, \cdots, N$. In this way, we artificially construct a conditional target distribution $p(\widetilde{\boldsymbol{t}}^{(n)}|\boldsymbol{s}^{(n)})$ that allows multiple plausible sentences conditioned on a particular source sentence.

Notice that, for the sake of constructing a distribution where the target sentences can have multiple plausible data points, there is no need to generate multiple random target sentences for a particular source sentence. In fact, it is preferred NOT, so that the experiment is more controlled. In the case where we generate a single target sentence $\widetilde{\boldsymbol{t}}^{(n)} = \boldsymbol{t}^{(\tau(n))}$ for a source sentence $\boldsymbol{s}^{(n)}$, $\{\widetilde{\boldsymbol{t}}^{(n)}|\boldsymbol{s}^{(n)}\}_{n=1}^N$ can still be viewed as samples from the marginal (unconditioned) distribution $p(\boldsymbol{t})$, and thus the desired "unaligned" property is in place.

It is straightforward to shuffle a subset of the translation dataset. This helps to analyze how Seq2Seq models behave when the "unaligned" problem becomes more severe. Shuffling trick is previously used by [11] to compare the robustness of Seq2Seq models and phrase-based statistical machine translation. Our paper contains a novel insight that shuffling datasets mimics the unaligned property in dialog datasets, which facilitates the comparison between Seq2Seq dialog and translation systems.

3.2. The Seq2Seq Model and Datasets

We adopted a Seq2Seq model (with an attention mechanism) as the neural network for both dialog and translation systems. The encoder is a bidirectional recurrent neural network with gated recurrent units (GRUs), whereas the decoder comprises two GRU transition blocks and an attention mechanism in between [12].[1]

For the dialog system, we used text-based dialog dataset, the Cornell Movie-Dialogs Corpus dataset,[2] containing 221k samples. For machine translation, we conducted experiments on a real-world dataset as well as a toy dataset. We applied the shuffling method in both two scenarios, mimicking the "unaligned" property. Following are the details of the translation datasets.

The Real-World Dataset. We used the WMT-2017 dataset[3] and focus on English-to-German translation, containing 5.8M samples. We trained Seq2Seq models on sub-word units by

[1] Code downloaded from `https://github.com/EdinburghNLP/nematus`

[2] Available at `https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html`

[3] Available at `http://data.statmt.org/wmt17`

Setting		BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Length	NLL	Entropy
Dialog	References	-	-	-	-	-	14.40	8.79	8.91
	Seq2Seq	1.84	15.1	2.40	1.02	0.66	11.70	8.08	7.92
Translation on WMT	References	-	-	-	-	-	21.47	11.4	10.2
	Seq2Seq	27.2	60.2	33.4	20.9	13.6	21.24	11.1	9.98
	shuffle 25%	24.4	56.2	30.3	18.8	12.0	21.02	10.9	9.81
	shuffle 50%	21.1	52.8	26.8	16.0	10.0	20.73	10.8	9.66
	shuffle 75%	17.2	48.2	23.2	13.4	8.10	19.89	10.6	9.39
	shuffle 100%	.024	12.5	.189	0.00	0.00	15.88	9.34	4.46
Translation on Toy	References	-	-	-	-	-	19.92	4.53	4.53
	Seq2Seq	99.9	99.9	99.9	99.9	99.9	19.92	4.53	4.53
	shuffle 25%	99.8	99.9	99.9	99.8	99.8	19.92	4.53	4.53
	shuffle 50%	40.2	51.9	41.2	40.1	40.0	17.34	4.21	2.31
	shuffle 75%	47.2	62.5	49.7	44.9	41.9	17.38	4.32	3.14
	shuffle 100%	.372	16.8	2.16	.181	.008	16.00	4.01	1.20

Table 1: BLEU scores, average length, negative log-likelihood (NLL), and entropy of dialog and translation systems.

using the Byte-Pair Encoding technique [13]. For validation and test, we used newstest2014 and newstest2016 sections, each containing 3k pairs respectively.

The Toy Dataset. We further evaluated our experiments on a toy dataset where we generated sequence-to-sequence samples from some predefined distributions. In this way, we can eliminate the effect of noise in real-world data (where source and target sides cannot match perfectly), serving as additional evidence of our claim.

In particular, the task for the toy dataset is to verbatim copy a source sequence. This can be thought of as a "trivial" translation dataset, where the source and target are exactly the same. Further, we sample the source string lengths and word frequencies from meaningful distributions so that the toy dataset more resembles true natural language.

Specifically, we first sample the length of source strings from a Poisson distribution, which is a counting distribution that oftentimes models the number of events in a certain time period. Formally, the probability of the length of a string being k is given by

$$p(\text{length} = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{1}$$

where λ is the parameter of the Poisson distribution, indicating the average length of strings, and in our case it was set to 20.

We then set the vocabulary to all lower-case letters in English. For simplicity, we do not model the dependency among characters in a string. In other words, each character is sampled from a unigram distribution. Considering that the frequency of words generally follows a power law distribution—also known as the Zipf's law [14]—in natural language, we also use a power law to approximate the unigram distribution. The power law has the form

$$p(x) \propto x^{\alpha-1}, \alpha > 1 \tag{2}$$

where α was set to 1.63 in our experiments. Our synthesized toy dataset consists of 500k training samples, 2k validation samples, and 2k test samples.

4. RESULTS

BLEU Scores. Table 1 presents the BLEU scores of dialog and machine translation systems. In open-domain dialog, BLEU-2 exhibits some (not large) correlation with human satisfaction [15], although BLEU scores are generally low. For machine translation, we achieved 27.2 BLEU for the normal setting on the English-to-German translation, which is comparable to 28.4 achieved by a baseline method in [16], and thus our replication of the machine translation system is fair. For the unshuffled toy dataset, we achieved 99.9 BLEU-4 score as expected, indicating that the aligned pattern is easy to be learned by the Seq2Seq.

If we begin to shuffle the WMT dataset, we see that BLEU drops gradually and finally reaches near zero if the training set is 100% shuffled. The results are not surprising and also reported in [11]. This provides a quick understanding on how the Seq2Seq is influenced by shuffled data.

The same phenomenon is also observed on the toy dataset. With the increase of the shuffling rate, the BLEU score on the toy dataset decreases.[4] The reason of this phenomenon is apparent. As the shuffling rate increases, "unaligned" problem becomes more and more severe, which makes the pattern in the dataset difficult to study. When the training data in the toy dataset is 100% shuffled, the BLEU score is close to zero while 50% and 75% shuffling settings return relative high scores, which is because simple alignment patterns still lies in the un-shuffled subset in the training data and are easy to be learned by the Seq2Seq model.

Length, Negative Log-Likelihood, and Entropy. We evaluated the quality of generated results (Table 1). The

[4]We regard the slight increase of BLEU score on the toy dataset from 50% to 75% shuffling as a result of randomness, which does not disprove our conjecture.

Setting		R^2 Correlation	
		Encoder	Decoder
Dialog	Seq2Seq	.5095	.1706
Translation on WMT	Seq2Seq	.9673	.8734
	+shuffle 25%	.9257	.7241
	+shuffle 50%	.9374	.6221
	+shuffle 75%	.8622	.6574
	+shuffle 100%	.4349	.8871

Table 2: R^2 correlation obtained by fitting a linear regression of the encoding/decoding step with hidden states.

length metric counts the number of words in a generated reply.[5] The *negative log-likelihood (NLL)* is computed as $-\frac{1}{|R|}\sum_{w\in R}\log p_{\text{train}}(w)$ where R denotes all replies and $p_{\text{train}}(\cdot)$ is the unigram distribution of words in the training set. *Entropy* is defined as $-\sum_{w\in R} p_{\text{gen}}(w)\log p_{\text{gen}}(w)$ where $p_{\text{gen}}(\cdot)$ is the unigram distribution in generated replies. Intuitively, both NLL and entropy evaluate how much "content" is contained in the replies. These metrics are used in previous work [4, 3], and are related to our research question.

We first compare the dialog system with machine translation, both in the un-shuffled setting. We observe that the dialog system does generate short and meaningless replies with lower length, NLL, and entropy metrics than references, as opposed to machine translation where Seq2Seq's generated sentences are comparable to references in terms of these statistics on both two datasets. Quantitatively, in the dialog system, the length is 20% shorter than references. The NLL and entropy decrease by 0.71 and 0.99, respectively; a decrease of 1 in NLL and entropy metrics is large because they are logarithmic metrics. Although with a well-engineered Seq2Seq model (with attention, beam search, etc.), the phenomenon is less severe than a vanilla Seq2Seq, it is still perceivable and worth investigating.

We then applied the shuffling setting to the translation system. With the increase of shuffling rate, the Seq2Seq model trained on translation datasets precisely exhibits the phenomenon as a dialog system: the length decreases, the NLL decreases, and the entropy decreases. In particular, the decreasing NLL implies that the generated words are more frequently appearing in the training set, whereas the decreasing entropy implies that the distribution of generated sentences spread less across the vocabulary. The phenomenon is consistent in both real and synthetic datasets.

In summary, artificially constructing an unaligned property in translation datasets—with all other settings remain unchanged—enables to reproduce the phenomenon in a dialog system. This shows evidence that the unaligned property could be one reason that causes the problem of short and meaningless replies in a dialog system.

Correlation between Time Steps and Hidden States. Shi et al. [17] conduct an empirical study analyzing "Why Neural Translations are the Right Length?" They observe that the length of generated reply is likely to be right regardless of the correctness of meaning. They further find that some dimensions in RNN states are responsible for memorizing the current length in the process of sequence generation; the result is also reported in [18]. Shi et al. [17] apply linear regression to predict the time step during sequence modeling based on hidden states, and compute the R^2 correlation as a quantitative measure.

Since a dialog system usually generates short replies (and thus not right length), we wonder whether such correlation exists in dialog and shuffled translation settings. We computed R^2 correlation as in [17] and show results in Table 2.[6] We find that the correlation is low with dialog system. In translation, the correlation first decreases then increases as the shuffling rate becomes larger. A possible explanation is that the lengths of generated translation sentences are similar when shuffling rate is at a high level.

5. CONCLUSION AND DISCUSSION

This paper addressed the question why dialog systems generate short and meaningless replies. We managed to reproduce this phenomenon in two translation datasets, artificially mimicking the scenario that a source sentence can have multiple equally plausible target sentences. Admittedly, it is impossible to construct identical scenario as dialog by using translation datasets (otherwise the translation just becomes dialog). However, the unaligned property is a salient difference, and by controlling this, we observe the desired phenomenon, demonstrating our conjecture.

Our findings also explain why referring to additional information—including dialog context [19], keywords [3] and knowledge bases [20]—helps dialog systems: the number of plausible target sentences decreases if the generation is conditioned on more information; this intuition is helpful for future development of text-based response generation in Seq2Seq dialog systems. Besides, our experiments suggest that Seq2Seq models are more suitable to applications where the source and target information is aligned.

6. ACKNOWLEDGMENTS

We thank all reviewers for their constructive comments. This research is supported by the National Key R&D Program under Grant No. 2017YFB1001804, and the National Natural Science Foundation of China under Grant No. 61832009.

[5] In some cases, an RNN fails to terminate by repeating a same word. Here, we assume a same word can be repeated at most four times.

[6] Because of the large noise of experiment results, the R^2 with 100% shuffling is the mean of 5 independent experiment results.

7. REFERENCES

[1] Lifeng Shang, Zhengdong Lu, and Hang Li, "Neural responding machine for short-text conversation," in *ACL (1)*. 2015, pp. 1577–1586, The Association for Computer Linguistics.

[2] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan, "A diversity-promoting objective function for neural conversation models," in *HLT-NAACL*. 2016, pp. 110–119, The Association for Computational Linguistics.

[3] Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin, "Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation," in *COLING*. 2016, pp. 3349–3358, ACL.

[4] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *AAAI*. 2017, pp. 3295–3301, AAAI Press.

[5] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, "Sequence to sequence learning with neural networks," in *NIPS*, 2014, pp. 3104–3112.

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

[7] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *CVPR*. 2015, pp. 3156–3164, IEEE Computer Society.

[8] Alexander M. Rush, Sumit Chopra, and Jason Weston, "A neural attention model for abstractive sentence summarization," in *EMNLP*. 2015, pp. 379–389, The Association for Computational Linguistics.

[9] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra, "Diverse beam search: Decoding diverse solutions from neural sequence models," *arXiv preprint arXiv:1610.02424*, 2016.

[10] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma, "Topic augmented neural response generation with a joint attention mechanism," *arXiv preprint arXiv:1606.08340*, 2016.

[11] Philipp Koehn, "Statistical machine translation (chapter 13: Neural machine translation)," *arXiv preprint arXiv:1709.07809*, 2017.

[12] Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde, "Nematus: a toolkit for neural machine translation," in *EACL (Software Demonstrations)*. 2017, pp. 65–68, Association for Computational Linguistics.

[13] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units," in *ACL (1)*. 2016, The Association for Computer Linguistics.

[14] George Kingsley Zipf, *The Psycho-Biology of Language*, Oxford, England: Houghton, Mifflin, 1935.

[15] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau, "How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *EMNLP*. 2016, pp. 2122–2132, The Association for Computational Linguistics.

[16] Antonio Valerio Miceli Barone, Jindrich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch, "Deep architectures for neural machine translation," in *WMT*. 2017, pp. 99–107, Association for Computational Linguistics.

[17] Xing Shi, Kevin Knight, and Deniz Yuret, "Why neural translations are the right length," in *EMNLP*. 2016, pp. 2278–2282, The Association for Computational Linguistics.

[18] Andrej Karpathy, Justin Johnson, and Li Fei-Fei, "Visualizing and understanding recurrent networks," *arXiv preprint arXiv:1506.02078*, 2015.

[19] Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao, "How to make context more useful? an empirical study on context-aware neural conversational models," in *ACL (2)*. 2017, pp. 231–236, Association for Computational Linguistics.

[20] Pavlos Vougiouklis, Jonathon S. Hare, and Elena Simperl, "A neural network approach for knowledge-driven response generation," in *COLING*. 2016, pp. 3370–3380, ACL.