LSDSCC: A Large Scale Domain-Specific Conversational Corpus for Response Generation with Diversity Oriented Evaluation Metrics

Zhen Xu[1], **Nan Jiang**[2], **Bingquan Liu**[1], **Wenge Rong**[2],
Bowen Wu[3], **Baoxun Wang**[3], **Zhuoran Wang**[3], **and Xiaolong Wang**[1]
[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
[3]Tricorn (Beijing) Technology Co., Ltd, Beijing, China
{zxu,liubq,wangxl}@insun.hit.edu.cn
{nanjiang, w.rong}@buaa.edu.cn
{wubowen, wangbaoxun, wangzhuoran}@trio.ai

Abstract

It has been proven that automatic conversational agents can be built up using the End-to-End Neural Response Generation (NRG) framework, and such a data-driven methodology requires a large number of dialog pairs for model training and reasonable evaluation metrics for testing. This paper proposes a Large Scale Domain-Specific Conversational Corpus (LSDSCC) composed of high-quality query-response pairs extracted from the domain-specific online forum, with thorough pre-processing and cleansing procedures. Also, a testing set, including multiple diverse responses annotated for each query, is constructed, and on this basis, the metrics for measuring the diversity of generated results are further presented. We evaluate the performances of neural dialog models with the widely applied diversity boosting strategies on the proposed dataset. The experimental results have shown that our proposed corpus can be taken as a new benchmark dataset for the NRG task, and the presented metrics are promising to guide the optimization of NRG models by quantifying the diversity of the generated responses reasonably.

1 Introduction

Conversational agents (a.k.a. Chat-bots) are effective media to establish communications with human beings and have received much attention from academic and industrial experts in recent years (Serban et al., 2017). One essential fact promoting the research work on conversational agents is the explosive growth of human interaction data accumulated in the social network services, such as Twitter[1] and Reddit[2]. So, it is possible to build Chat-bots based on data-driven approaches (Serban and Pineau, 2015).

[1]https://twitter.com/
[2]https://www.reddit.com/

Nevertheless, there still remains a great challenge for building such conversational agents: at present, the automatic evaluation metrics of NRG models can hardly afford to measure the semantic relevance and diversity of generated results reasonably, and even the latter evaluation aspect has been paid little attention. The widely accepted evaluating methods employed by the existing NRG models can be categorized as: a) metrics inherited from Machine Translation, e.g., BLEU, Perplexity, etc. (Yao et al., 2015; Lowe et al., 2017; Wu et al., 2018); b) discrete scores measuring the quality of generated results by human labeling (Shang et al., 2015; Serban et al., 2016; Xu et al., 2017); and c) case study comparing the generated results of different NRG models (Shang et al., 2015; Wang et al., 2017). The disappointing situation is that these evaluating methods have not revealed tangible difference among NRG models, the reasons for which can be reflected by the example given in Table 1.

Query: Where did you get that from?
Ground-truth responses: I got it from her.
– I do not know. – Cloverfield wiki. – New York Times. – From movie theatre.
Query: Airplane is now available on Netflix!
Ground-truth responses: Thank you!
– Is it worth watching? – Thank you for that I'll add it to my list. – Awesome, I haven't watched it!

Table 1: Cases of queries with diverse responses.

For each query in Table 1, one response from the testing set is taken as the ground truth, together with responses with more morphological and semantic variations, marked with the symbol "–". These samples indicate that the numerical metrics inherited from NMT which discard the diver-

Proceedings of NAACL-HLT 2018, pages 2070–2080
New Orleans, Louisiana, June 1 - 6, 2018.

sity among responses, cannot reflect marginal differences among generative models, which is supported by the research work of Liu et al. (2016). Thus, an NRG model with good capability to produce diverse and meaningful responses is possible to be judged as a poor one by the BLEU/Perplexity based evaluations. Meanwhile, the metrics based on human labeling are still promising, yet the expensive cost and inconsistency among labelers limit the scale of human-annotation. Therefore, it becomes a necessity to develop reasonable automatic evaluation metrics, that can be taken to measure both candidate response's diversity and its relevance to the given query, to effectively guide the training of NRG models towards the state of promoting meaningful and diverse responses (Li et al., 2016a; Shao et al., 2017; Freitag and Al-Onaizan, 2017).

In order to evaluate the performance of NRG models automatically and reasonably, a well-annotated testing set should be built first. But then, building such a high-quality testing set is a non-trivial task indeed. On one hand, most existing source datasets cover various domains, making it difficult to evaluate the generated results in case that the domain of the generated response is different from that of the reference. On the other hand, a large number of noises, typos, and slangs distribute in existing large-scale datasets, such as Twitter corpus (Ritter et al., 2011) and Ubuntu dialog corpus (Kadlec et al., 2015). For instance, there are many file directories with computer names in Ubuntu dialog corpus. Therefore, qualified domain-specific datasets are desperately required to evaluate NRG models with different architectures reasonably.

To address the above issues, we build a high-quality and domain-specific dialog corpus composed of a carefully prepared training set, and meanwhile, a testing set is constructed by collecting multiple reference responses for each query and conducting group-aware human annotation on collected responses. On this basis, we proposed three discriminative metrics: `MaxBLEU`, Mean Diversity Score (`MDS`), and Probabilistic Diversity Score (`PDS`), to primarily evaluate the diversity of generated responses with relevance also considered. To further assess the performance and effectiveness of the test set cooperating with the proposed metrics, the widely applied Sequence-to-Sequence (Seq2Seq) (Bahdanau et al., 2014; Sutskever et al., 2014) based models with the available diversity promotion methods are implemented, and experiments are conducted on the proposed Large Scale Domain-Specific Conversational Corpus (LSDSCC) dataset. The experimental results stay consistent with the previous experience acquired from human-labeled sets, and the performance of these models suggests that the LSDSCC corpus and discriminative metrics will provide insights for future research in the field of NRG.

2 Related Work

Seq2Seq based conversation modeling approaches have been proven to be able to generate response directly (Vinyals and Le, 2015; Shang et al., 2015). However, these models tend to produce generic responses to any given queries, namely the *deficient diversity problem* (Shao et al., 2017). Recent studies attempt to constrain these universal replies and promote more diverse responses with various strategies during the procedure of training or inference (Li et al., 2016a,b; Mou et al., 2016; Xing et al., 2017; Shao et al., 2017). Besides, there still exits another meaningful option, that is, to employ reasonable diversity oriented evaluation metrics to guide the optimization of models.

The quality of testing sets is a primary factor for such evaluation of NRG models. Existing large-scale corpora contain the Movie Dialogue, Ubuntu, Twitter, and Reddit corpus (Banchs, 2012; Uthus and Aha, 2013; Ritter et al., 2010; Schrading et al., 2015). The Ubuntu corpus is built by scraping a large scale tech support dialogues from Ubuntu IRC forum for building response ranking models (Kadlec et al., 2015). Similarly, Sordoni et al. (2015) provide external context information for message response pairs from Twitter FireHose. Besides, Dodge et al. (2016) and Schrading et al. (2015) collect real conversations from movie categories of Reddit community, which are integrated into a multi-task corpus on movie for the ranking task and discourse analysis. In the above corpora, there are only one or two reference responses for most query, which is completely unlike that of the practical conversation scenario (Li et al., 2017b). By contrast, this paper construct a high-quality testing set, including multi-references for each query. In this regard, our testing set is more close to the real-world setting.

Besides the testing set, evaluation metrics are also important for the performance measurement of NRG models. Most frequently applied evaluation metrics for NRG models are inherited from NMT to measure the fluency and relevance of generated responses, such as Perplexity, BLEU (Papineni et al., 2002) and deltaBLEU (Galley et al., 2015). Although these metrics demonstrate the relevance between the given query and the generated responses, they overlook the reply's diversity that is of great importance in conversation setting. Thus, efforts are devoted to simulate the human subjective judgment, which is similar with the response ranking task in retrieval-based chat agents (Lowe et al., 2017; Tao et al., 2018), but unavoidable uncertainty and errors are brought into the systems (Hu et al., 2014). In addition, automatic evaluation metrics (e.g. BLEU, deltaBLEU, etc.) are limited by the fact that each query only has references with the exact same meaning and many overlapped phrases, which is unreasonable in the conversational scenario.

3 Data Processing and Analysis

Previous studies indicate that more focused topics and less diverged domain is helpful to guide NRG models away from the state of producing universal responses (Mou et al., 2016; Xing et al., 2017), so we compose a domain-specific corpora by constraining the domain of crawled dialogues from Reddit to its movie discussion board[3]. The quality of the data in Reddit movie category has been discussed by Stoddard (2015) and Jamnik and Lane (2017), who point out that the popularity is a good indication of relative quality and the movie category is one of the most popular boards in Reddit. Thus, the data in Reddit movie category is originally high-quality. In this section, the pipeline for building the LSDSCC dataset will be discussed in detail, and necessary statistical indicators are collected to demonstrate its distribution. Moreover, human evaluation is conducted to measure the quality of the obtained training set.

3.1 Data Processing

Data Cleansing. We crawl threads from the movie discussion board of Reddit that includes human-to-human conversations as the raw dataset, and conduct the following cleansing operations:

a) For each thread, we strip away the *markdown* and *html* syntax tokens, e.g., "[word](url)" is transformed to "word", ">" is reformed to ">", etc. Meanwhile, all forms of urls, emails and digits withing the paragraphs are normalized as "url", "email" and "digits" tokens respectively;

b) As emoticons in the data originated from social media services always provide essential emotional information of users, we propose to convert the same groups of emoticons into corresponding words (e.g., ":-)" will be reverted to "happy") to preserve such emotion knowledge;

c) Finally, replicated words or characters (e.g., "cooool" and "ahahaha", etc.) are substituted with its normal form using regular expressions.

Query Vocabulary	Size	Coverage (%)
overlong words	17,084	10.32 %
non-ascii words	58,720	35.46 %
Response Vocabulary	**Size**	**Coverage (%)**
overlong words	15,914	7.94 %
non-ascii words	71,997	35.94 %

Table 2: Composition of noise words in the query and response vocabulary of the raw data.

Vocabulary Truncation. After the above preprocessing operations, there still exist redundant unformatted slang and noisy strings (e.g., "Iloveyou"), which have low-frequency in the crawled raw data. Consequently, the vocabulary size of the dataset is exceeding $160K$ as shown in Table 2. Keeping a such large vocabulary for Seq2Seq based models will consume excessive memory and make those models difficult to converge, while pruning low-frequency unformatted slang and noisy strings into "UNK" symbols would directly harm the performance of the model since excessive knowledge hidden in these strings are ignored in the training process. To address this issue, we break these slangs and noises into several frequent words in our corpus and eliminate non-ASCII tokens. In this way, sufficient information of the dataset is maintained for model training. Finally, the vocabulary sizes of the dataset are reduced to around $50K$.

Dialog Pruning. Statistical results on the sentence length of query-response pairs in the cleaned corpus are illustrated in Fig. 1. Concerning the fact that recurrent neural networks can not efficiently capture the semantics of over-long sentences and previous studies indicate that such re-

[3]https://www.reddit.com/r/movies/, selected from https://www.reddit.com/r/datasets

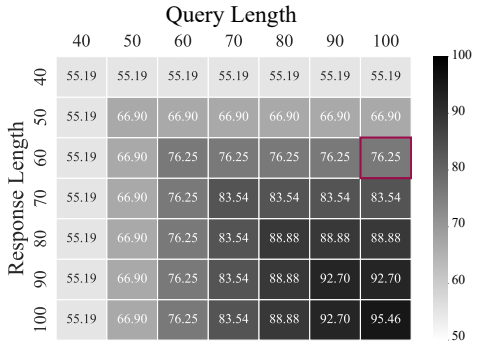

Figure 1: Sentence length coverage of queries and responses within the dataset.

sponses would make the decoder hard to converge (Greff et al., 2017), it is necessary to prune the pairs containing very long sentences. After the sentences are tokenized using the *NLTK toolkits*[4], the cases with queries longer than 100 or responses exceeding 60 words are pruned directly, and 76.25% of the dataset are finally reserved.

After pruning the corpus, there remain 738,095 single-turn and 346,543 multi-turn conversations. Since this paper focuses on the single-turn dialogs, the evaluative testing set and detailed experiments in the following sections are designed for single-turn corpus. As the testing set will select pairs from the preprocessed data, the corresponding tuples will be deleted to avoid coverage.

3.2 Query-Response Relevance of the Dataset

As one of the most important qualities of the conversational corpus, the query-response relevance demonstrates the overall quality of the dataset. Human evaluations of the query-response relevance are conducted to validate the quality of the dataset used in this paper. Nine experienced annotators are invited to evaluate the query-response relevance of 500 single-turn dialogs uniformly sampled from the whole dataset obtained in Subsection 3.1. In the evaluation, we ask each annotator to label whether the response is appropriate to the corresponding query in the given query-response pair. A pair is tagged as "Unsure" if the annotator could not confirm the degrees of relevance without related context and background movie knowledge. The labeled result is shown in Table 3. It is observed that 85% samples in the query-response relevance task are confirmed to keep high relevance between the query and the corresponding responses. Moreover, there exist only about 6.6% irrelevant noises. So, the resource can be considered as a high-quality one and can be used in the practical task.

Category	Relevant	Unsure	Irrelevant
Numbers	427	40	33
Percentage	85.4%	8%	6.6%

Table 3: Query-Response Relevance on the single-turn training set.

4 Testing Set and Evaluation Metrics

Existing evaluation metrics of dialog agents measure the quality of the generated sentences only by referring to the existing responses, which obeys the same principle with NMT models' metrics. However, one essential difference between NRG and NMT lies in the fact that, a large group of responses can be considered as relevant to a given query in conversations, while the number of references to a translation result is quite limited for NMT models. So the diversity degree of candidates which have not covered by NMT oriented evaluation metrics, is supposed to be quantified and measured in NRG models.

Currently, few studies focus on the evaluation based on the group of references, which is more meaningful and reasonable for NRG models. Therefore, we proposed three metrics: `MaxBLEU`, Mean Diversity Score, and Probabilistic Diversity Score, to quantify both the relevance and diversity of the generated responses. Since these metrics are based on the multi-reference, we first describe the procedure of building testing set, with multi-references for each query. Then, the metrics for NRG models are detailed based on the multi-reference testing set.

4.1 Multi-Reference Testing Set Construction

Fig. 2 illustrates the response quantity distribution of queries in the preprocessed data. While the testing set is randomly sampled from the preprocessed data, the response quantity distribution of the testing set is the same as that in Fig. 2. In this case, the multi-reference testing set for NRG evaluation is difficult to construct by directly extracting samples from the dialog corpus, since there are too few queries that contain more than three responses. Roughly choosing samples from such data is possible to bring topic bias into the testing set, and manually filtering suitable candidate pairs from

[4]http://www.nltk.org/

them is also time-consuming and expensive. Nevertheless, there exist large amounts of queries that are highly semantically similar or correlated with each other. This indicates that the multiple references can be obtained by selecting responses of queries that are semantically identical to the original query. What's more, the human-annotation is involved to proofread the filtered pairs' quality and complete the final labeling.

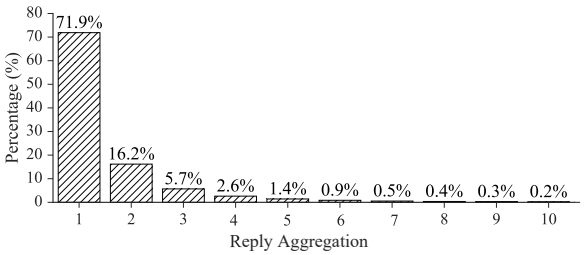

Figure 2: Distribution of reply quantities in training set.

When constructing the testing set, the very first step is getting semantically similar (or even identical) queries with the given ones. For this purpose, this paper adopts the TF-IDF similarity and semantic embedding based distance to measure the similarity between queries. The procedure of gaining similar queries is divided into two stage: In the first stage, we employ Apache Lucene[5] to exploit the word-level TF-IDF patterns within queries, and then extract the top 100 similar queries with highest scores given by Lucene for each query. Yet, these candidates only capture n-gram level similarity with the probably diverged semantics. Thus, in the second stage, we utilize paragraph vector algorithm (a.k.a Doc2vec[6]) (Le and Mikolov, 2014) to resort the selected similar queries in the semantic space and only queries of similarity score higher than a certain threshold (i.e., 0.9) are reserved. Table 4 lists several identical queries filtered by Lucene and Doc2vec methods with the given query. It should be noted that the Lucene index and Doc2vec need to be initialized by feeding all the sentences in the dialogue corpus.

To reserve as much information as possible and balance the distribution of the composed testing set, we divide the dataset into several subsets based on the response number of queries, and then sample testing data from each subset uniformly. Concretely, according to the response number, queries of the dataset are divided into three subsets: a) queries with less than 3 responses, b) queries with 3 to 5 responses, and c) queries with more than 5 responses. We randomly sample 100 queries from each subset, and thus 300 queries are obtained as the testing set. Aiming at building a multiple references testing set, each query in the testing set is assigned with 15 responses, including the original responses and the ones of the most similar queries obtained by the procedure of last paragraph.

Similar Queries	Similarity
If you haven't already, watch the animatrix.	0.97
Do not watch the animatrix, you may leave you house.	0.95
I don't have much to ad except, that people really should watch animatrix.	0.94
I recommend you watch the matrix.	0.91

Table 4: Filtered queries identical to the original query: "**You should watch the animatrix.**"

Afterwards, three skilled and experienced labelers familiar with movies are employed and carefully trained to crosswise annotate the filtered testing set. In addition, labelers can also obtain some background of the corresponding query since there are additional details for most queries in Reddit. In this case, the quality of selected samples can be guaranteed. Besides, the annotators are asked not only to label the relevance of query and reference responses, but also reorganize the independent references into groups by the semantic similarity subjectively. The grouping strategy is introduced for the purpose of evaluating the diversity of responses generated by different models.

In the relevance oriented annotation procedure, the labelers are first asked to judge whether a candidate response is appropriate and natural to the input query. If a candidate response is grammatically correct and semantically relevant with the corresponding query from the annotators' perspective, it should be labeled as "**1**". Otherwise, the annotators have to give "**0**" label to the candidate. Then, for each query, the annotators need to split responses labeled with "**1**" into different groups based on word overlapping between them, with stop-word overlapping ignored. Finally, the groups with the similar semantics are merged into a larger group by the annotators, so as to get the final grouped responses.

[5]https://lucene.apache.org/

[6]https://radimrehurek.com/gensim/models/doc2vec.html

At last, we obtain a high-quality testing set, in which each query is assigned with different numbers of reference responses. Fig. 3 shows the distribution of the response numbers in the testing set. Comparing to the original response number distribution in Fig. 2, the replies distribution of the testing set is much more appropriate for the conversational scenario. Furthermore, responses to the corresponding query are categorized into several groups. In this case, NRG models can be evaluated reasonably using such a testing set. One sampled case in the testing set is shown in the left phase of Fig. 4, and there are eight responses in the labeled data divided into four groups. The different metrics in this figure will be introduced in the following sections. It should be noted that both the single-turn dialogs and the annotated testing set are released[7].

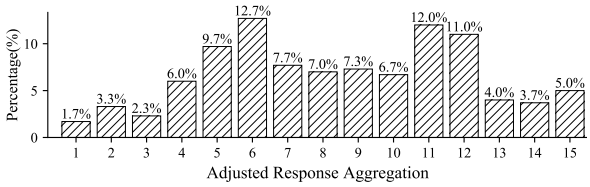

Figure 3: Aggregated responses per query in testing set.

4.2 The Metric on Response Relevance

Since the NRG architecture is analogous to the NMT models, introducing the BLEU scores to evaluate the semantic relevance of the generated results is acceptable. However, it is not reasonable to average the BLEU scores of the generated response to each reference, because the semantic of each reference varies significantly. Aiming at revealing the variation and diversity among responses, which have not yet covered at the NMT models, we propose a `MaxBLEU` metric customized for response generation based on the Multi-BLEU metric (Madnani et al., 2008). Noticing that the metrics inherited from SMT, like BLEU, is not able to evaluate the diversity of responses, we propose the specified metrics for diversity evaluation, which will be described in the next subsection.

Given an input query, the NRG model is able to generate a set of hypothesis $\{h_i\}$[8]. Meanwhile, according to the human-annotation strategy described in Subsection 4.1, the set of references can be reorganized, based on their semantic similarity, into the groups with the format of $\{r_{ij}\}$, where r_{ij} denotes the j-th reference in the i-th group. On this basis, the `MaxBLEU` metric is defined as:

$$\texttt{MaxBLEU}(h_i) = \arg\max_k \text{Multi-BLEU}(h_i, r_{k\cdot}) \quad (1)$$

where $r_{k\cdot}$ denotes all the references in the k-th group. That is, we begin by calculating all the multi-BLEU scores between each hypothesis and grouped references, and pick the score for the sentence with the highest BLEU as the score for this set of hypothesis, so that we make an alignment between generated hypothesis h to the group-aware references r. For simplicity, one response can only be aligned to one group reference, and multi-group references are not considered in this work.

4.3 Metrics on Response Diversity

Given a query, the diversity degree of candidate responses is an essential criterion for evaluating the performances of NRG models. Currently, most studies tend to demonstrate the diversities of different models by sampling and comparing the generated results, or labeling the diversity of the generated samples, which makes it difficult to benchmark and automatically evaluate different models. Although Li et al. (2017a) propose to calculate the number of distinct unigrams and bigrams of generated responses, such scores do not align well with human inspection (Serban et al., 2017).

Algorithm 1 Two Response Diversity Metrics.

Input:
hypothesis set `H` and reference set `R` ;
Output:
Mean Diversity Score (`MDS`);
Probabilistic Diversity Score (`PDS`);
for all $r_{i\cdot} \in$ `R` **do** ▷ Initialize
  $p_i = 1/|\texttt{R}|$,
  $p'_i = |r_{i\cdot}|/\sum_j |r_{j\cdot}|$.
end for
for all $h_i \in$ `H` **do** ▷ Compute alignment
  $k = \arg\max_j$ Multi-BLEU$(h_i, r_{j\cdot})$.
  $\texttt{MDS}(k) = p_k$,
  $\texttt{PDS}(k) = p'_k$.
end for
return $\sum_k \texttt{MDS}(k), \sum_k \texttt{PDS}(k)$

Therefore, we propose two evaluative metrics based on the `MaxBLEU` metric for diversity measurement: a) Mean Diversity Score (`MDS`) and b) Probabilistic Diversity Score (`PDS`). Basically, the

[7] https://drive.google.com/file/d/1nbpbnhwNP14xAc4SAc1-NN5lvEr01dQb/view?usp=sharing

[8] Following the terms in machine translation, this part takes "hypothesis" to represent "response".

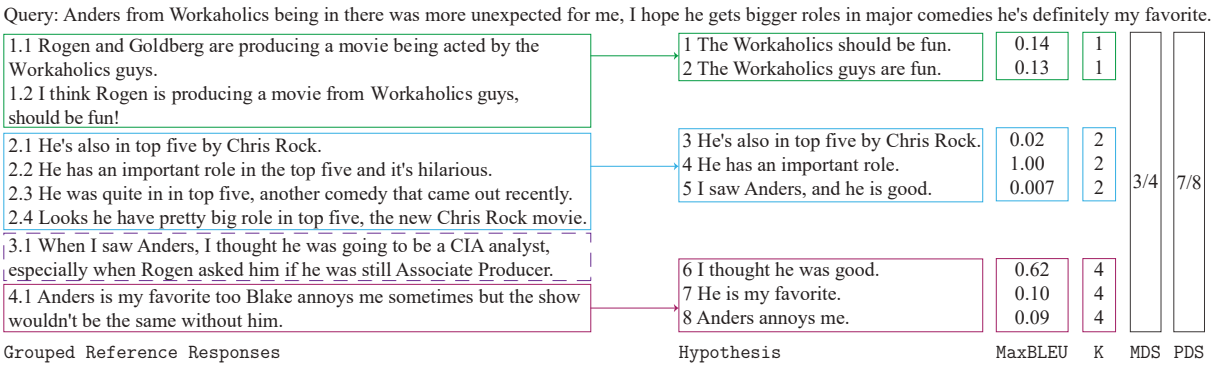

Figure 4: A sampled testing case including a given query, the grouped reference responses and the generated ones (hypothesis) with the proposed metrics, among which "K" is obtained by performing $\arg\max$ upon MaxBLEU scores. The MDS metric is calculated with the partitioned groups and PDS metric is calculated with the weight of each group in the overall the candidate set.

two metrics aim at measuring the overall diversity of the whole set of generation results (hypothesis) by taking them as an entirety, and the detailed calculation steps of the proposed metrics are illustrated in Algorithm 1. According to the algorithm, the PDS metric assumes that the weight of each reference group is distributed uniformly, regardless of the reference number in each group. Similarly, the MDS metric takes the count of the members in each group as the weight of the corresponding group, and actually compute the weighted coverage upon the reference group.

5 Experiments and Analysis

In this section, we present the detailed experiments on the single-turn dialog dataset and analysis on generated results, in accordance to the proposed metrics.

5.1 Baselines and Experimental Setups

Experiments are conducted using the popular Seq2Seq based models with the currently available diversity prompting strategies as follows:

1) **Basic Seq2Seq.** We employ the basic Seq2Seq to build the encoder-decoder architecture running on the proposed dataset, by taking the bidirectional LSTM cell as the encoder to address the input sentences ordering problem and classic LSTM cell as the decoder (Vinyals and Le, 2015).

2) **Attention-Seq2Seq.** As proposed by Vinyals and Le (2015); Luong et al. (2015), a concatenated version of attention mechanism is applied upon the basic Seq2Seq model.

3) **Greedy-Seq2Seq.** Based on the basis Seq2Seq model, the diversity promotion strategy proposed by Li et al. (2016b) is applied in the generating procedure, and the training procedure stays the same. Hyper parameter γ, a.k.a. *diversity rate*, are set with empirical experiments (i.e., $\gamma = 0.1, 0.8$) to reveal the efforts.

4) **Greedy-Attn-Seq2Seq.** Following the work of Li et al. (2016b), the greedy diversity promotion strategy is applied on the Seq2Seq model with attention mechanism similar with model 2, and we set hyper parameter $\gamma = 0.1, 0.8$.

5) **MMI-Seq2Seq.** In the generation procedure, Maximum Mutual Information (MMI) model is applied in the decoder to prune generic answers on the basic Seq2Seq model (Li et al., 2016a).

In our research, we implement these models on the TensorFlow platform[9], and Adam optimizer (Kingma and Ba, 2015) is employed for gradient optimization during training. Besides, we choose to prune the words whose frequencies are below 2, so the source and target vocabulary are set to $42,257$ and $46,865$ respectively.

In addition, we set the batch size to 50, hidden size of encoder to 256, hidden size of decoder to 512 and learning rate to $2e-4$. The gradients are clipped within $[-3.0, 3.0]$ to avoid the gradient explosion problem. Every model runs on a single GPU separately for at least one week before convergence. Afterwards, for all these methods, we generate a set of hypothesis sentences with beam size set to $k = 50$, and the evaluation scores are obtained using the proposed metrics.

After running through 25 epochs on the dataset, the training log-loss of the basic Seq2Seq mod-

[9]https://www.tensorflow.org

Models	MaxBLEU	MDS	PDS
Seq2Seq (Vinyals and Le, 2015)	1.30	0.230	0.253
Attention-Seq2Seq (Luong et al., 2015)	1.42	0.235	0.262
Greedy-Seq2Seq ($\gamma = 0.1$) (Li et al., 2016b)	1.88	0.249	0.243
Greedy-Seq2Seq ($\gamma = 0.8$)	1.72	0.297	0.291
Greedy-Attn-Seq2Seq ($\gamma = 0.1$) (Li et al., 2016b)	**2.27**	0.252	0.248
Greedy-Attn-Seq2Seq ($\gamma = 0.8$)	2.05	0.285	0.287
MMI-Seq2Seq (Li et al., 2016a)	2.15	**0.311**	**0.329**

Table 5: Performances of different models trained on the LSDSCC dataset with three metrics: MaxBLEU, MDS, PDS.

els converge to about 4.2 and the Seq2Seq models augmented with attention converge to 3.1. Also, we set the dropout rate to 0.5, which enables us to tune the models though much more epochs and avoid the over-fitting problems.

5.2 Relevance Analysis

The semantic relevance of the generated responses is represented by the MaxBLEU scores, which are listed in the corresponding column of Table 5. From this benchmarking table, it can be observed that the attention mechanism is helpful for decoders to improve the relevance of the generated responses, since the *Attention-Seq2Seq* performs better than the basic Seq2Seq on the dataset, in terms of all the three metrics. However, the relative gain of the attention layer is limited, indicating that modeling relation of query and response by attention module is not able to directly solve the learning paradigm of conversations.

In accordance to the results of Greedy-Seq2Seq ($\gamma = 0.1$) and Greedy-Seq2Seq ($\gamma = 0.8$), the hyper-parameter γ actually plays an important role in the generation steps of the decoder. Since γ is introduced to constrain the selection probability of the next-step word by performing the re-ranking process, and the larger value of this parameter will lead to the greater impact upon generating steps and produce more diverse sentences, we evaluate this greedy strategy with γ set with two empirical value. It can be seen that the model with the smaller γ performs better than the one with the larger parameter, which can be attributed to the fact that responses with more diversity are less similar to references. Similar observation can be get from the results of models Greedy-Attn-Seq2Seq ($\gamma = 0.1$) and Greedy-Attn-Seq2Seq ($\gamma = 0.8$). Besides, the reason for setting $\gamma = 0.1, 0.8$ in this part is that they are well represented for the poor diversity and good diversity, which the exact score of γ will vary under different configurations and structures of model.

In addition, the MMI model is proved to be promising to enhance the generation models, by improving both the relevant and diversity of generated responses. Even though the *MMI-Seq2Seq* model has not got the highest MaxBLEU, it outperforms the other ones on diversity, which will be discussed in the following subsection.

5.3 Diversity Analysis

Table 5 also illustrates the MDS and PDS score of each benchmark. It is observed that the greedy strategies in the generating procedure with the greater parameter γ can boost the diversity of generated responses obviously. This phenomenon is attributed to the inter-sibling ranking policy in the decoding procedure, which tends to choose hypotheses from diverse parents. In addition, the MMI strategy gets the highest MDS and PDS, because the MMI criterion relieves the constraint of the language model, under which general responses always get a higher generative probability.

Meanwhile, the PDS metric aligns well with the basic MDS, but the relative gap becomes larger within the Greedy-Seq2Seq ($\gamma = 0.1$) and Greedy-Seq2Seq ($\gamma = 0.8$) models. The reason for enlarging relative gap between different models, is to distinguish the performance of similar models and evaluate the performance of specific module inside the models. When comparing Seq2Seq and Attention-Seq2Seq, relative gain of applied attention module to the overall model in terms of MDS was 2.1%, while it became 3.6% considering the PDS metric.

Practically, it is reasonable to make a trade-off between the relevance and diversity. The PDS is more suitable for choosing the systems with stringent diversity requirement, and the MDS is a softer metric, which should be taken into consideration when measuring the diversity improvements by integrating some new modules into NRG models.

Moreover, it can be observed that the relevance oriented metric MaxBLEU gets improvement along

with the increasing of the diversity oriented PDS and MDS. This phenomenon indicates a relationship between relevance and diversity against that in some of text generation tasks (e.g., image caption (Yao et al., 2017)). Since there are generally many references for a given query, the relevance and diversity are possible to be improved simultaneously for the response generation task (Li et al., 2016a). And thus, the topic changing on the generated results is tolerable.

5.4 Human Correlation Analysis

To validate the correlations between human ratings and the proposed metrics, we further invite 9 annotators with rich movie knowledge to judge the relevance and diversity of the generated responses from benchmark methods. Each baseline model generate 10 responses for each query in the test dataset. The annotators are first asked to judge whether a generated response is relevant to the query (labeled with 1) or not (labeled with 0). After that, the annotators estimate the diversity of relevant responses of each query with a scale of 1 to 3. The final Fleiss Kappa (Fleiss, 1971) score is 0.46, which denotes moderate agreement of the annotators.

Model	Spearman	p-value	Pearson	p-value
MaxBLEU	0.31	0.022	0.29	0.036
MDS	0.36	0.041	0.33	0.028
PDS	0.39	0.038	0.35	0.040

Table 6: Correlation between the proposed metrics and human judgments for the Reddit dataset.

The Pearson and Spearman correlation between the human evaluations and each metric are given in Table 6. It can be observed that the proposed metrics correlate with human judgments moderately with $p - value < 0.05$, which is quite different from the correlation test in Liu et al. (2016). This can be attributed to the fact that there are multiple references for each query in our test dataset. Although the proposed metrics are derived from the word-overlap based BLEU scores, expanding references of each query makes such scores much more reasonable for evaluating the relevance and diversity of generated responses.

6 Conclusion and Future Work

In this paper, we have proposed the Large Scale Domain-Specific Conversational Corpus (LSDSCC), collected from the movie discuss threads in the Reddit community, for training and testing the Neural Response Generation (NRG) models. In addition, necessary data cleansing and pruning works are done to remove noises in the utterances. Moreover, we employ volunteers to annotate a diverse query-responses testing set, with reference groups taken into consideration for objectively quantifying the diversity of generated results. On the basis of the testing set, we propose two evaluative diversity metrics (mean diversity score and probabilistic diversity score) calculated according to the standard MaxBLEU score.

Furthermore, we investigate the performance of popular Seq2Seq based models with various diversity promotion strategies, and the score of them are collected to validate the effectiveness of the proposed metrics. The proposed dataset and evaluation metrics are expected to be used for the effective training and reasonable testing of NRG models.

In the future studies, we would explore the possibility of promoting diversity on the learning procedure, by directly optimizing diversity loss in the cost function. Besides, injecting external information during response's generation would be another challenging work.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. This research is partially supported by National Natural Science Foundation of China (No.61572151, No.61602131, No.61672192) and the National High Technology Research and Development Program ("863" Program) of China (No.2015AA015405).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*.

Rafael E. Banchs. 2012. Movie-dic: a movie dialogue corpus for research and development. In *Proc. of ACL*, pages 203–207.

Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. 2016. Evaluating prerequisite qualities for learning end-to-end dialog systems. In *Proc. of ICLR*.

J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proc. of ACL*, pages 445–450.

K. Greff, R. K. Srivastava, J. Koutnk, B. R. Steunebrink, and J. Schmidhuber. 2017. Lstm: A search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.*, 28(10):2222–2232.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Proc. of NIPS*, pages 2042–2050.

Matthew R. Jamnik and David J. Lane. 2017. The use of reddit as an inexpensive source for high-quality data. *Practical Assessment Research & Evaluation*, 22(5).

Rudolf Kadlec, Martin Schmid, and Jan Kleindienst. 2015. Improved deep learning baselines for ubuntu corpus dialogs. In *Machine Learning for SLU Interaction Workshop, NIPS*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proc. of ICML*, pages 1188–1196.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proc. of NAACL-HLT*, pages 110–119.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. A simple, fast diverse decoding algorithm for neural generation. *CoRR*, abs/1611.08562.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017a. Learning to decode for future success. *CoRR*, abs/1701.06549.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017b. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proc. of IJCNLP*, pages 986–995.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proc. of EMNLP*, pages 2122–2132.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proc. of ACL*, pages 1116–1126.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proc. of EMNLP*, pages 1412–1421.

Nitin Madnani, Philip Resnik, Bonnie J Dorr, and Richard Schwartz. 2008. Are multiple reference translations necessary? investigating the value of paraphrased reference translations in parameter optimization. In *Proc. of AMTA*, pages 143–152.

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. In *Proc. of COLING*, pages 3349–3358.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Proc. of NAACL-HLT*, pages 172–180.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proc. of EMNLP*, pages 583–593.

Nicolas Schrading, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on reddit. In *Proc. of EMNLP*, pages 2577–2583.

Iulian V. Serban and Joelle Pineau. 2015. Text-based speaker identification for multi-participant open-domain dialogue systems. In *Machine Learning for Spoken Language Understanding and Interaction, NIPS 2015 Workshop*.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2017. A survey of available corpora for building data-driven dialogue systems. *CoRR*, abs/1703.05742.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proc. of AAAI*, pages 3776–3784.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proc. of ACL*, pages 1577–1586.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017.

Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proc. of EMNLP*, pages 2210–2219.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proc. of HLT-NAACL*, pages 196–205.

Greg Stoddard. 2015. Popularity dynamics and intrinsic quality in reddit and hacker news. In *ICWSM*, pages 416–425.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proc. of NIPS*, pages 3104–3112.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proc. of AAAI*, pages 1–8.

David C. Uthus and David W. Aha. 2013. The ubuntu chat corpus for multiparticipant chat analysis. In *AAAI Spring Symposium: Analyzing Microtext*.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. In *Proc. of ICLR*.

Jianan Wang, Xin Wang, Fang Li, Zhen Xu, Zhuoran Wang, and Baoxun Wang. 2017. Group linguistic bias aware neural response generation. In *Proceedings of the 9th SIGHAN Workshop on Chinese Language Processing*, pages 1–10.

Yu Wu, Wei Wu, Dejian Yang, Can Xu, Zhoujun Li, and Ming Zhou. 2018. Neural response generation with dynamic vocabularies. In *Proc. of AAAI*.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proc. of AAAI*, pages 3351–3357.

Zhen Xu, Bingquan Liu, Baoxun Wang, SUN Chengjie, Xiaolong Wang, Zhuoran Wang, and Chao Qi. 2017. Neural response generation via gan with an approximate embedding layer. In *Proc. of EMNLP*, pages 617–626.

Kaisheng Yao, Geoffrey Zweig, and Baolin Peng. 2015. Attention with intention for a neural network conversation model. In *Machine Learning for SLU Interaction Workshop, NIPS*, pages 1–7.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2017. Incorporating copying mechanism in image captioning for learning novel objects. In *Proc. of CVPR*, pages 5263–5271. IEEE.