The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Towards a Neural Conversation Model with Diversity Net Using Determinantal Point Processes

Yiping Song,[1] Rui Yan,[2,3*] Yansong Feng,[2] Yaoyuan Zhang,[2] Dongyan Zhao,[2,3] Ming Zhang[1*]
[1]Institute of Network Computing and Information Systems, School of EECS, Peking University, China
[2]Institute of Computer Science and Technology, Peking University, China
[3]Beijing Institute of Big Data Research, China
{songyiping,ruiyan,fengyansong,zhang_yaoyuan,zhaody,mzhang_cs}@pku.edu.cn

Abstract

Typically, neural conversation systems generate replies based on the sequence-to-sequence (seq2seq) model. seq2seq tends to produce safe and universal replies, which suffers from the lack of diversity and information. Determinantal Point Processes (DPPs) is a probabilistic model defined on item sets, which can select the items with good diversity and quality. In this paper, we investigate the diversity issue in two different aspects, namely query-level and system-level diversity. We propose a novel framework which organically combines seq2seq model with Determinantal Point Processes (DPPs). The new framework achieves high quality in generated reply and significantly improves the diversity among them. Experiments show that our model achieves the best performance among various baselines in terms of both quality and diversity.

Introduction

Automatic conversation systems, facilitating a smooth interaction in natural languages between humans and computers, are of growing importance in both academia and industry. Recently, with massive publicly accessible free-chatting resources on the Web and the fast development of data-driven deep learning techniques, it becomes more and more promising for us to build a non-task-oriented conversation system.

Deep learning has greatly advanced neural conversation systems. Given a human utterance, called a *query*, a neural conversation system generates a *reply* tailored for the query. Most neural conversation systems are based on the sequence-to-sequence (seq2seq) model (Sordoni et al. 2015; Shang, Lu, and Li 2015; Serban et al. 2016b), which is derived from neural machine translation. seq2seq model encodes a query into a vector (also known as *encoder*), and decodes the vector into a reply to the query (also known as *decoder*). This framework has been widely used in neural conversation systems due to its strong capability of capturing the semantic relevance between queries and replies.

However, seq2seq model aims at finding the most relevant "translation" sentence during the decoding process, which is not applicable to the conversation scenario. Given a query, there are multiple choices to respond it: replies can be completely different but all appropriate. Although a traditional seq2seq model can produce multiple reply candidates during beam search, the most top-ranked replies from the beam search have minor differences with each other (Li et al. 2016a). As showed in Table 1, compared with the replies in various expressions from human-beings, providing multiple similar replies to a given query all the time would make users feel boring.

Query	*Human* Reply	*System* Reply
Nice weather today, isn't it?	Yeah, can't be better!	The weather is good!
	Great weather for outdoor activities!	The weather is great!
	But it is going to rain tomorrow.	Good weather!
I know the new movie will be showed soon.	I heard it would be next month.	I don't know.
You know what, Bob is coming!	Let's get together for a little party?	I don't know.
How to get to the Great Wall?	Sorry, this is my first time here.	Sorry, I don't know.

Table 1: An example of how humans respond versus how traditional generative systems respond. A human can offer different proper replies to the same query. On the contrary, the traditional systems could only generate similar replies, and they prefer to use universal replies for different queries.

Moreover, due to the origin of the seq2seq model, the model "translates" the inputs with the maximum likelihood as the outputs. Universal replies, such as "I don't know," "I'm OK", seem to be plausible for most queries and have a dominant coverage in natural conversation datasets.[1] For a traditional seq2seq model, it is safe to "translate" most queries into such universal replies with the maximum likelihood. Some cases are in Table 1, providing a universal reply (which is actually meaningless) to many queries, the conversation system becomes boring.

As discussed, we point out two aspects of diversity in neural conversation systems: (1) *query-level* diversity and (2) *system-level* diversity. The former means the inner-query diversity. For a particular query, we aim to generate different replies with different semantics to respond the query appropriately. The latter means the outer-query diversity. For a conversation system, we manage to solve the problem to generate universal replies for various user queries.

Different from some previous works, which incorporate

*Corresponding authors

[1]According to Li et al. (2016a), 0.45% sentences contain the sequence "I don't know."

the context (Serban et al. 2016b; Tian et al. 2017; Yan, Song, and Wu. 2016; Yan et al. 2016), keyword (Mou et al. 2016; Yao et al. 2017) or knowledge-base (Yin et al. 2016) into the reply generation process, we do not use additional information but foster diversity by making full potential of seq2seq model.

In this paper, we propose to connect the single-turn neural conversation system using the seq2seq model to determinantal point processes (DPPs) (Kulesza, Taskar, and others 2012) as a joint generation framework. DPPs define probability measures on sets and the maximum a posteriori (MAP) decoding algorithms (Yao et al. 2016) could make use of that measures for selecting the items with both good quality and diversity. This property fits the demands of generative conversation systems as it is effective to produce both meaningful and diverse words and replies. In our master model. we first formulate diversity in every word choosing during the beam search via a diversity net, then employ a DPP decoding strategy to reorder the subsequences for the next generation state. Besides the master model, we also propose a simplified but still effective re-ranking model for easier implementation. In the experiments, we examine the performance of both model variants against several baselines, and experimental results indicate the effectiveness of our proposed models.

To sum up, our contributions are as follows:

• We systematically investigate two different aspects of diversity, i.e., *query-level* diversity and *system-level* diversity, and tackle them simultaneously through a unified framework.

• We connect the seq2seq model to Determinantal Point Processes (DPPs) in a neural conversation system to achieve both quality and diversity in the generated replies, which is a new insight. We propose two model variants to incorporate diversity *in* and *after* beam search.

Preliminaries

We introduce the traditional seq2seq model and Determinantal Point Processes (DPPs) as preliminaries.

seq2seq Model

seq2seq is a prevailing model, which is first introduced in statistical machine translation to transform one language into another. Now, the conversational generation is treated as a monolingual translation task (Ritter, Cherry, and Dolan 2011; Shang, Lu, and Li 2015), which translates the user-issued query to an appropriate reply.

The seq2seq model consists of two parts, namely the encoder and the decoder. The encoder maps the user-issued query Q into a distributed vector and the decoder uses this vector to generate the corresponding reply R. Both the encoder and decoder apply recurrent neural networks (RNNs) to model sentences. To further improve the ability of RNNs, gating mechanism such as gated recurrent unit (GRU) (Cho et al. 2014) and long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997) are used to improve the quality of longer sentences. Attention mechanisms (Bahdanau, Cho, and Bengio 2014) are used to strengthen the connection between the encoder and the decoder.

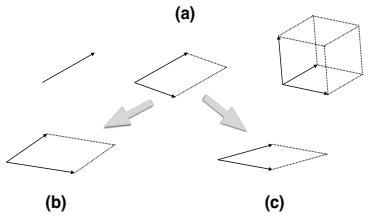

Figure 1: The geometric interpretation of DPPs. (a) The probability of a subset Y is the square of the volume spanned by its corresponding item's feature vectors, where each feature vector describes the property an item. (b) The probability of a subset Y increases when the magnitude of the item's feature increases. (c) The probability of a subset decreases when the similarity between two corresponding items increases.

The standard objective function for the seq2seq model is the log-likelihood of reply R given the query Q (Shang, Lu, and Li 2015). Thus the replies with high probabilities are more likely to be generated, which explains why seq2seq tends to generate universal replies.

DPPs

The determinant point process is a probabilistic model defined on item sets. Based on this probabilistic model, many efficient ms could be performed to solve the problems such as sampling, marginalization, conditioning and other inference tasks (Kulesza and Taskar 2010). The main application of DPPs is to select items as diverse as possible and guarantee the quality of selected ones at the same time. It has been applied to document summarization (Gillenwater, Kulesza, and Taskar 2012), image search (Kulesza and Taskar 2010) and clustering (Kang 2013).

Given a candidate set $\mathcal{Y} = \{c_1, c_2, \cdots, c_n\}$, a positive semidefinite L called *L-ensemble* depicts the probability of each subset $Y \subseteq \mathcal{Y}$:

$$p(Y) = \frac{\det(L_Y)}{\det(L+I)} \tag{1}$$

where I is the identity matrix of the corresponding dimension and Y is a random subset of $\mathcal{Y}$. $L_Y \equiv [L_{i,j}]_{i,j \in Y}$ denotes the restriction of L to the entries indexed by elements of Y, and $\det(L_\emptyset) = 1$.

There is an intuitive geometric interpretation of DPPs. If L is written as a Gram matrix, such that $L = B^\top B$ (B can always be found since L is positive semidefinite). Denoting the columns of B by B_i for $i = 1, 2, \ldots, n$, where:

$$p(Y) \propto \det(L_Y) = \mathrm{Vol}^2\left(\{B_i\}_{i \in Y}\right) \tag{2}$$

where Vol refers to the $|Y|$-dimensional volume of the parallelepiped spanned by the columns of B corresponding to elements in Y (Figure 1). Details could be seen in (Kulesza, Taskar, and others 2012), which is beyond the scope of this paper.

Methodology

We propose two model variants adopting DPP to incorporate diversity *in* and *after* generation and give the explicit measurement of quality and diversity, which is essential for DPP modeling.

Model 1: DPP Decoder (DPP-D)

Under the framework of `seq2seq`, beam search is a prevalent method for approximate decoding (Li et al. 2016a; Li and Jurafsky 2016; Vijayakumar et al. 2016). However, these replies obtained from beam search are only a poor surrogate for the entire search base (Huang 2008; Finkel, Manning, and Ng 2006), which causes some replies overlap with each other closely; the minor differences usually lie in punctuations and different tenses. In the standard beam search, there remains a set of top-k (known as the *width* in beam search) subsequences. using the scoring function given by

$$\begin{aligned} score(y_1, \ldots, y_t) = \log p(y_t|Q, y_1, \ldots, y_{t-1}) \\ + score(y_1, \ldots, y_{t-1}) \end{aligned} \tag{3}$$

where t is the index of iteration. In the right side of the equal sign, the first part is the probability of each word in the entire vocabulary, and the second part is the score of the subsequence passed from the last iteration. Since the value of second part is accumulated as the length of subsequences grows, it could easily exceed the value of the first part and dominate the score, which causes the problem the high scored subsequences are mostly from the same or similar ancestors.

To boost the diversity of generated results, we propose DPP-D model, which undertakes the encoder part of `seq2seq` model for query representation and promotes the vanilla decoder in two ways. One is to use diversity net for word choosing at each time step, and another is to adapt DPPs to re-rank the corresponding subsequences of choose words from diversity net. Specifically, in each time step, the model keeps k candidate subsequences for further generation, as same as the standard beam search. The pipeline of decoding is comprised of three components, namely GRU cell, Diversity net and DPP selecting.

GRU cell. GRU cell takes the word from the last time step as input, and uses standard GRU equations to update the hidden state, then outputs a vector representing the probability distribution over the entire vocabulary. In the traditional `seq2seq` model, the probability is utilized to choose the top-k ranked ones for the next time step. As discussed before, this distribution assigns similar words with similar probabilities, which impairs the diversity among chosen words. In stead of traditional beam search, we keep $3k$ top-ranked words and give the further ranking task to the diversity net, where it could extract words with good diversity.

Diversity Net. Diversity net takeovers the hidden state from GRU cell and outputs $2k$ proper words. Diversity net employs DPP to locate the more important and diverse neural nodes in networks. Let $\{v_1^l, \cdots, v_t^l, \cdots, v_m^l\}$ and $\{v_1^{l-1}, \cdots, v_p^{l-1}, \cdots, v_q^{l-1}, \cdots, v_n^{l-1}\}$ be the neural nodes in layer l and $l-1$ respectively, where v_i^l donates the i-th node in layer l. $\boldsymbol{v}_i^l$ is the representing vector of v_i^l, and it summarizes the activation from last layer $l-1$: $\boldsymbol{v}_i^l = (a_{i,1}^l v_1^{l-1}, a_{i,2}^l v_2^{l-1}, \cdots, a_{i,n}^l v_n^{l-1})$, where $a_{i,j}^l$ is the parameters between layer l and $l-1$ (Mariet and Sra 2016). In this way, each node in layer l has a vector to describe its property. Based on these vectors, we perform DPP to find the important and diverse nodes in layer l . Only the selected nodes would be sent to the next layer and perform further nodes selection. At each layer l, we first create an $m \times m$ matrix L by setting $L_{i,j}$ as:

$$L_{i,j} = L_{j,i} = \|\boldsymbol{v}_i^l - \boldsymbol{v}_j^l\|^2 \tag{4}$$

Then, the selecting process finds the best subset of good quality and diversity from the whole node set. This is equivalent to finding the set $Y \subseteq \mathcal{Y}$ that maximizes $\det(L_Y)$, which has been proved to be an NP-hard problem (Gillenwater et al. 2014). Here, $\mathcal{Y}$ is the whole node set in layer l, Y is a subset of $\mathcal{Y}$. In this paper, we select the highest probability node set via an approximate greedy approach named MAP decoding (Yao et al. 2016). In this algorithm, the score for a subset S_Y is defined as the unnormalized log probability given L : $score_L(Y) = \log\det(L_Y)$. This algorithm (showed in Algorithm 2) has formal approximation guarantee for the worst cases and runs very fast in practice (Yao et al. 2016). After this, the nodes in the chosen subset Y are regarded as the activated nodes in the current layer, and only the activated nodes would be passed to the next layer for further calculation.

We use diversity net to find $2k$ diverse neural nodes in the top layer of GRU, which calculates a probability distribution over entire vocabulary. Due to the merit of DPP, nodes in the chosen subset are the ones with good quality and diversity. Since each chosen nodes corresponds to a word and a subsequence stored during the generation process, these subsequences would be passed to the next selecting step.

DPP Selecting. Since we obtain $2k$ candidate words from diversity net, resulting in $2k$ subsequence. DPP selecting reorders the $2k$ candidates and pass the top-k for further generation.

We first evaluate the subsequences as the preparation for further selection. We establish the probabilistic matrix L on the subsequence set where L could be written as $L = B^{\mathrm{T}}B$. We regard the columns of B as feature vectors describing each subsequence. Thus L can be defined as:

$$L_{i,j} = q_i \phi_i^{\mathrm{T}} \phi_j q_j \tag{5}$$

where q_i is the *quality* of the subsequence i, and ϕ_i is the feature vector of subsequence i, so the $\phi_i^{\mathrm{T}}\phi_j$ measures the *similarity* between subsequence i and subsequence j. We propose the *quality* measurement $\mathrm{Qua}(\cdot)$ to calculate the quality score q_i. $\phi_i^{\mathrm{T}}\phi_j$ refers to the similarity between two subsequences, which is used to measure the diversity among the whole set. We define the *diversity* measurement $\mathrm{Div}(\cdot)$ to calculate $\phi_i^{\mathrm{T}}\phi_j$. These two measurements would be explained in Section . Then, we perform the same DPP selection algorithm described above.

The procedure of DPP-D is showed in Algorithm 1. Figure 2 presents the main idea of DPP-D.

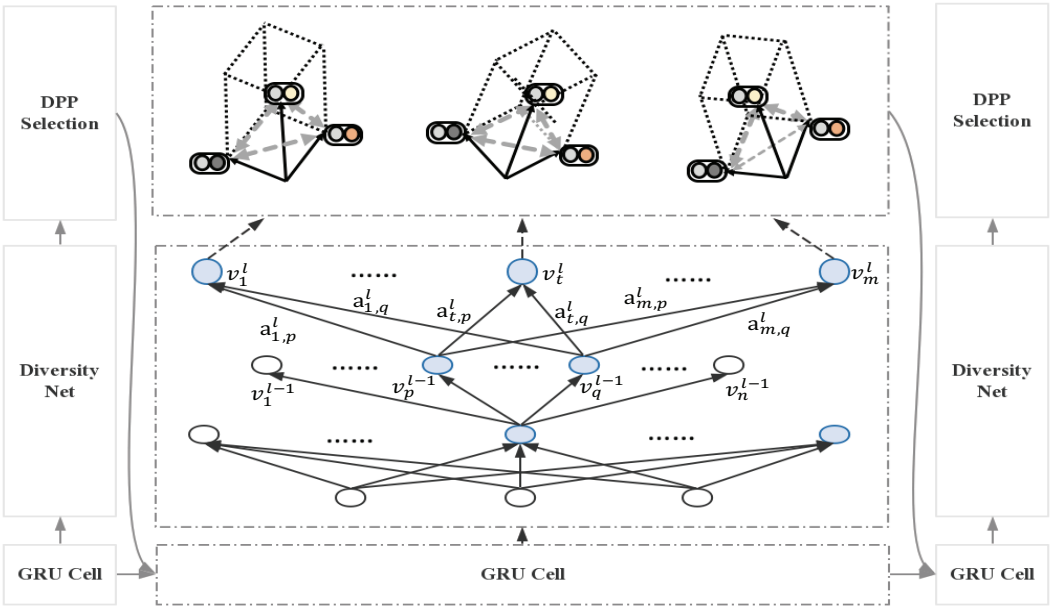

Figure 2: The decoder part of DPP-D model.

Model 2: DPP Re-ranker (DPP-R)

As the DPP-D model requires continuous iterations of decoding, we design a simplified model **DPP-R** by taking the decoding and DPP selection as separate components.

Concretely, DPP-R first performs standard beam search to obtain a list of reply candidate set containing $k * k$-best replies, known as the beam search process. Then DPPs are used to upturn the potential of the k-best replies which is the DPP re-ranking process. Each candidate reply is measured by both quality and diversity to build the L matrix. After all those preparation, we perform the greedy DPP MAP inference (Yao et al. 2016) (showed in Algorithm 2) to obtain the re-ranked reply set. At each time step, the algorithm greedily chooses the current best reply, and it terminates when we obtain enough chosen replies.

Thus DPP-R reranks all the candidate replies obtained from beam search after generation, and the newly ranked candidate set upturns these replies that are of good quality and diversity.

Quality and Diversity Measurement

To perform DPPs involved selections, an explicit definition of both the *quality* and *diversity* is needed. In this section, we introduce the quality and diversity measurements.

Quality Measurement. For single-turn conversation systems, the coherence between the user-issued query and each candidate reply is used to evaluate the goodness of candidate replies. In this paper, we define the quality of each candidate c (reply or subsequence) by semantic coherence between Q and c, which is based on word-level similarity. For each word in the query, we find the best matching word in the candidates using the cosine similarity of word embeddings, then average all the cosine similarity scores as the final quality measurement, given by

$$\mathrm{Qua}(c, Q) = \frac{1}{|Q|} \sum_{w_j \in Q} \underset{w_i \in c}{\operatorname{argmax}} \cos(\boldsymbol{e}_{w_i}, \boldsymbol{e}_{w_j})$$

where Q is the query, c is the candidate, $|Q|$ is the number of the words in Q, and $\boldsymbol{e}_{w_i}$ is the embedding of the word w_i.

Diversity Measurement. As described in the previous section, diversity of a set is indirectly characterized with a similarity in a DPP model. Hence, the diversity measurement composes of two similarity scores: 1) term overlap score, 2) embedding matching score. The term overlapping score represents each candidate c_i and c_j as one-hot word vector, and measures the similarity by the cosine function; it explicitly reveals the word overlap between candidates. The embedding matching score uses the cosine similarity of word embeddings; it captures the underlying semantic matching between two candidates. We linearly combine these two features as the final diversity score, that is,

$$\mathrm{Div}(c_i, c_j) = \lambda \frac{1}{|c_j|} \sum_{w_{jl} \in c_j} \underset{w_{ik} \in c_i}{\operatorname{argmax}} \cos(\boldsymbol{e}_{w_{jl}}, \boldsymbol{e}_{w_{ik}}) + (1 - \lambda) \cos(\boldsymbol{h}_i, \boldsymbol{h}_j)$$

where $\boldsymbol{h}_i$ is the one-hot word vector of the c_i, each element indicating if a word appears in c_i, $|c_i|$ is the number of word in c_i, λ is the hyperparameter balancing the two similarities.

Experiments

Experimental Setups

We evaluated our approach on a massive Chinese conversation dataset crawled from Baidu Tieba.[2] There were

[2]http://tieba.baidu.com

Algorithm 1: The DPP-D algorithm

Input: User-issued query Q
Maximum length of each reply l
Beam search width k
Output: Generated reply set $\mathcal{R}$
//Repeat alternatively to compete the generation process
$S_0 = \varnothing$ // S_t is the subsequence set at time t
for $t = 1; t \leq l;$ **do**
  $V_t^{'} = \text{GRU\_Cell}(S_{t-1}, Q)$
  // $V_t^{'}$ contains $3k$ words
  $S_t^{'} = \text{Diversity\_Net}(V_t{'})$
  // $S_t^{'}$ contains $2k$ subsequences
  // DPP Selection
  $L_{|S_t'| \times |S_t'|} = \{0\}$ // L is the *L-ensemnble*
  for $s_i \in S_t^{'}$ **do**
    $q_i = Qua(s_i, Q)$
    for $s_j \in S_t^{'}$ & $j \leq i$ **do**
      $\phi_i^{\mathrm{T}} \phi_j = \text{Div}(s_i, s_j)$
    $L_{i,j} = q_i \phi_i^{\mathrm{T}} \phi_j q_j$
    $L_{i,j} = L_{j,i}$
  S_t = DPP_selection(L, k)
  // S_t contains k subsequences
$\mathcal{R} = S_l$
return $\mathcal{R}$

Algorithm 2: DPP selection

Input: L, number of items to be chose k
Output: Selected set
$S = \varnothing$
repeat
  $s = \text{argmax}_{s \in S_t'} \log\det L(S \cup \{s^{'}\})$
  $S = S \cup s$
until $|S| = k$;
return $\mathcal{S}$

1,600,000 query-reply pairs for training, 2000 pairs for validation, and another unseen 2000 pairs for testing. We also performed standard Chinese word segmentation.

All the models are under the architecture of traditional `seq2seq` model[3]. All the proposed models were implemented during the testing, following Li and Jurafsky (2016). We used the bi-directional recurrent neural network with gated recurrent units (Bi-GRU RNN) (Serban et al. 2016a) to capture the information along the word sequences. To train the neural conversation models, we followed the hyperparameter settings in (Shang, Lu, and Li 2015; Song et al. 2016). The word embeddings were 610d and hidden layers were 1000d. We applied AdaDelta with default hyperparameters, where batch size is 80. We kept 100K words (Chinese terms) for queries, and 30K for replies due to efficiency concerns. The beam size k was 20, so the proposed methods would not suffer from the efficiency problem as only 60 ($3k$) nodes would be sent to diversity net. We kept the top-10 generated replies for each query in the diversity evaluation. Notice that the validation set was used for an early stop based on the perplexity measurement. The word embeddings used in the *quality* measurement was pre-trained on 3 million utterances with 150k unique words. The hyperparameter λ used in *diversity* measurement was empirically set to 0.8.

[3]Codes and sample data will be soon available at: https://github.com/stellasyp/DPP-Conversational-System

Competing Methods

We compared our models with the vanilla beam search and other various diversity enhanced models. All the methods are trained in the same way to guarantee a fair comparison.

• **Standard Beam Search (SBS).** It is the standard decoding method used in `seq2seq` model.

• **Diverse Decoding (DD).** DD improves the standard beam search by punishing the bottom ranked subsequences among siblings (Li and Jurafsky 2016) [4].

• **Diverse Beam Search (DBS).** DBS divides the subsequences into several groups during decoding and assigns lower scores to the groups which are similar to the prior groups (Vijayakumar et al. 2016).

• **DPP-R.** Our DPP-R re-ranks the k-best replies via DPPs selection, which incorporates diversity *after* decoding.

• **DPP-D.** Our DPP-D applies DPPs at every step of the decoding process, which incorporates diversity *in* decoding. This one does not include the diversity net strategy.

• **DPP-D-DivNet.** Our DPP-D-DivNet applies DPPs at every step of the decoding process with diversity net, which incorporates diversity *in* decoding. This one is the master model of this paper.

Evaluation Metrics

We evaluated the quality and the diversity of the generated replies obtained from each method in both subjective and objective manners. We implemented the objective evaluation on the whole testing set, and randomly sampled 100 cases for the subjective evaluation which is time- and labor-consuming.

Quality Evaluation

• **BLEU Score.** BLEU-1 and BLEU-2 are used as the automatic evaluation (Papineni et al. 2002), which are correlation-related metrics used in conversation systems (Li et al. 2016b; Mou et al. 2016). We calculated the BLEU scores of top-1 reply to assess the performance of each method and display the average BLEU scores of top-10 replies.

• **Human Annotation.** We asked 3 annotators to label the quality of the top-1 replies (Mou et al. 2016). Each reply would be labeled as "0" for bad, "1" for borderline and "2"

[4]A similar work by Li et al. (2016a) has been demonstrated to be less effective than DD, so we did not include it.

for good. This evaluation was conducted in a strictly random and blind fashion to rule out human bias.

Diversity Evaluation

• **Distinct Score and Diversity Score.** For each competing method, we calculated the *system-level* diversity and *query-level* diversity. The *system-level* diversity measures the different expressions among the whole set of queries. We calculated the number of distinct 1-, 2-, 3- and 4-grams in all the generated replies, which are known as the distinct scores for diversity measurement (Li and Jurafsky 2016; Vijayakumar et al. 2016).

The *query-level* diversity measures the diversity of the different expressions within one particular query. Since we kept top-10 candidate replies for each query, the inner-query diversity scores were calculated by the accumulated similarities between these 10 sentences. The diversity score function is computed as:

$$\frac{2}{|C|(|C|-1)} \sum_{c_i \in C} \sum_{c_j \in C \setminus \{c_i\}} (1 - \cos(\boldsymbol{h}_i, \boldsymbol{h}_j)) \qquad (6)$$

C is the candidate set, $\cos(\cdot)$ is the cosine measure and $\boldsymbol{h}_i$ is a one-hot vector (each element indicating if a word appears in c_i). This function is used in Zhang and Hurley (2008) to evaluate the diversity of a set.

• **Human Annotation.** We invited 3 educated annotators to rate the query-level diversity and systems-level diversity. The query-level diversity indicates the percentage of generated sentences which have a different meaning from any other candidates among all top-10 generated sentences (noted as *Clustering*).

To evaluate the systems-level diversity, we asked volunteers to label the universal replies, such as I dont know, Im OK, in all top-1 replies to calculate the percentage of universal replies (noted as *Percentage*). We did not define the standards of the different and univerisal expressions but let the annotators to use their own understanding.

Performance and Analysis

We present the quality performance in Table 2 and diversity performance in Table 3. Figure 3 presents a sample of annotated cases (3 annotators are agreed on this sample).

In Table 2, SBS shows a relatively poor performance. DD is slightly better than SBS in terms of BLEU scores and human evaluation. DBS achieves the worst performance in terms of BLEU scores and even not as good as the SBS, indicating this method may pay too much attention to improve the diversity but ignore the quality of generated replies. DPP-R is better than all the methods above, which means it is simple and effective. DPP-D does not use diversity net but applies DPP selecting strategy, and it still achieves a significant improvement of all metrics. Comparing with DPP-R, we conduct the conclusion that integrating diversity "in" each step of decoding is a better choice than "after" it. Our master model DPP-D-DivNet achieves the best performance among all the competing methods and outperforms the others by a large margin. The fact that DPP-D-DivNet is prior to DPP-D demonstrates that diversity net indeed benefits our model. The BLEU scores of the top-10 replies and human score have a similar result. It is natural that the top-10 replies which cover a larger set of candidates may be a better way to measure the replies' quality. Various expressions of replies could be suitable to respond the particular query in conversation systems, but there is only one reference for the calculation of BLEU scores. Hence, it makes more sense to have better BLEU scores for top-10 results than top-1 results.

	Utterance (Translated)	Meaning
Query	我9号到拉萨，求组队。微信：xxxxxx (I will arrive in Lhasa on 9th. Join me and let's team up for this trip! WeChat: xxxxxx)	–
SBS	OK (OK.) QQ: xxxxx (QQ: xxxxx) 微信：xxxxxx (WeChat: xxxxxx) 我加你微信吧 (I will add you on WeChat.) 我加你微信聊吧 (I will add you and talk with you on WeChat.)	1 2
DD	OK .(OK.) QQ: xxxxx (QQ: xxxxx) 我2号到拉萨 (I will arrive in Lhasa on 2nd.) 加微信吧：xxxxxx (Add me on WeChat: xxxxxx) 加微信：xxxxxx(Add me, WeChat: xxxxxx)	1 2 3
DBS	OK .(OK.) 加群UNK (Add our chatting group UNK) 具体时间呢？ (What is the specific time?) 可以的啊. (OK.) 我们7号到丽江 (We will arrive in Lijiang on 7th)	1 2 3 4
DPP-R	我6号到拉萨 (I will arrive in Lhasa on 6th.) 你加我微信: xxxxxx (Add my WeChat: xxxxxx) 什么时候出发？ (When will you leave?) 几号出发？ (What date will you leave?) 你们在拉萨么？ (Are you in Lhasa?)	1 2 3 4
DivNet DPP-D	我们7号出发 (We will leave on 7th) 求同行，求组队! (Let's go together, let's team up!) 要去尼泊尔么？ (Will you go to Nepal?) 具体出发时间还没定呢 (The exact departure time is not decided.) 加我QQ吧：xxxxxx (Add me on QQ: xxxxxx)	1 2 3 4 5

Figure 3: Examples of the top 5 replies. Numbers in gray indicate different meanings. QQ and WeChat are two popular chatting tools. "xxxxxx" is the numbers omitted for privacy. "UNK" is the out-of-vocabulary word.

In Table 3, it can be seen that SBS is the worst method since it does not consider the diversity. DBS is better than the DD and DPP-R in most cases but is slightly worse than DPP-D. DPP-R performs well in diversity and even beats DPP-D in terms of *distinct-1* metric, which indicates that the use of DPP to rerank the k-best replies sometimes is also a good strategy (although not always). DPP-D almost excels all the methods above, that means the dpp selection is an effective strategy. DPP-D-DivNet has the best performance both in *system-level* and *query-level* diversity, either in subjective or in objective evaluations. In the human annotation experiments, 8.7 out of 10 replies have distinct meanings and only 5.333% generated replies are universal replies.

By jointly analysing the experimental results in two tables, we could see that DD is not as good as DBS, which is also showed in Vijayakumar et al. (2016). DBS, which shows a relatively good performance in diversity, does not have a satisfactory performance in quality. So we may assume that DBS sacrifices the quality of replies to increase the diversity. DPP-R makes good use of the original beam search and greatly improves the diversity. DPP-D model achieves the best performance both in quality and diversity.

Method	Top-1 Reply		Top-10 Replies		Human Score
	BLEU-1	BLEU-2	BLEU-1	BLEU-2	
SBS	1.053	0.420	3.711	1.520	0.590
DD	1.160	0.380	4.108	1.484	0.737
DBS	0.363	0.078	3.144	0.790	0.747
DPP-R	2.698	1.399	4.897	1.926	0.613
DPP-D	3.157	0.897	7.827	2.278	0.927
DPP-D-DivNet	**8.568**	**3.740**	**9.914**	**16.56**	**1.440**

Table 2: **Quality** measurement in terms of BLEUs and average human scores. Inter-annotator agreement for human annotation: Fleiss' $\kappa = 0.5077$ (Fleiss 1971), std $= 0.3791$, indicating moderate agreement (within a proper range in (Mou et al. 2016)).

Method	*system-level* Diversity					*query-level* Diversity	
	distinct-1	distinct-2	distinct-3	distinct-4	Percentage	Diversity	Clustering
SBS	0.004	0.023	0.049	0.076	55.33%	0.867	2.877
DD	0.005	0.029	0.061	0.092	40.67%	0.907	3.851
DBS	0.020	0.167	0.384	0.547	15.67%	0.938	6.748
DPP-R	0.026	0.128	0.249	0.358	30.67%	0.920	6.392
DPP-D	0.023	0.171	0.387	0.572	7.333%	0.952	8.317
DPP-D-DivNet	**0.058**	**0.327**	**0.619**	**0.809**	**5.333%**	**0.956**	**8.727**

Table 3: **Diversity** measurement in terms of *system-level* diversity and *query-level* diversity. Inter-annotator agreement for human annotation: for **Percentage** , Fleiss' $\kappa = 0.9138$, std $= 0.0603$; for **Clustering** Fleiss' $\kappa = 0.4993$, std $= 0.5325$.

Related Work

Conversational Systems. Automatic conversation system is a prolonged research topic in natural language processing (NLP). In early days, most of conversation systems served for vertical domains. Now, conversation systems in the open domain have become the center of attention. Retrieval paradigm and Generation paradigm are two main approaches in the open domain (Song et al. 2016). With the increasing development of neural networks, generative conversation systems demonstrate powerful capabilities to learn from human dialogue patterns.

Diversity in Conversation Systems. A hot research topic in generative conversation systems is the generated replies are lack of diversity. Several works are focusing on the improvement of diversity. Li et al. (2016a) abandon the standard objective function in `seq2seq` model, and replace it with maximum mutual information as the training criterion. Li and Jurafsky (2016) propose a diverse decoding method based on the traditional `seq2seq`, which avoids the sibling sequences deriving from the same ancestral sequences during the decoding process.[5] Another alternative idea is to divide the candidate sequences into several groups, and assign lower scores for the groups that are much similar to other groups (Vijayakumar et al. 2016). Shao et al. (2017) propose a target-glimpse model with a fixed-length decoder with self attention. These methods improve the diversity of replies, but may ignore the coherence between queries and replies. Besides, Zhao et al boost diversity via conditional variation autoencoder (Zhao, Zhao, and Eskenazi 2017), but this work regards chatting history as the condition, which does not fit for single-turn conversation scenario.

Determinantal Point Processes. Determinantal point procese is applicable in selection problems such as image search (Kulesza and Taskar 2010) and recommender systems (Gillenwater et al. 2014) where a small number of items with good quality and diversity are chosen from a much larger set. Various probabilistic inference approaches can be performed efficiently, including sampling, marginalization and conditioning (Gillenwater, Kulesza, and Taskar 2012; Kulesza, Taskar, and others 2012). Vondrák, Chekuri, and Zenklusen (2011) depicts a greedy algorithm which could choose the best item in every selection process. Kulesza and Taskar (2011) propose to sample a certain number of items with the highest probability from the whole candidate set. The most relevant approach is a diversity net (Mariet and Sra 2016) for model compression but this one is trainable.

[5]This work is addressed in the scenarios of the machine translation, which could be naturally adapted to conversation systems.

Conclusion

In this work, we investigate the diversity issue in neural conversation systems from two aspects: *query-level* and *system-level* diversity. To tackle this problem, we propose to connect `seq2seq` model with the determinantal point processes for jointly modeling the quality and diversity of the generated replies. We come up with two various DPP-based generative models namely DPP-D and DPP-R. The experiment results show that our proposed methods could achieve the best performance against the existing state-of-the-art models.

Acknowledgement

We thank Jin-ge Yao and Lili Mou for their inspiring discussions. This paper is partially supported by the National Natural Science Foundation of China (NSFC Grant Nos.61772039, 61472006, 91646202 and 71672058) and the National Key Research and Development Program of China (No.2017YFC0804001). Rui Yan was sponsored by CCF-Tencent Open Research Fund.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. In *Computer Science*.

Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Finkel, J. R.; Manning, C. D.; and Ng, A. Y. 2006. Solving the problem of cascading errors: Approximate bayesian inference for linguistic annotation pipelines. In *EMNLP*, 618–626.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. In *Psychological Bulletin*, volume 76, 378. American Psychological Association.

Gillenwater, J. A.; Kulesza, A.; Fox, E.; and Taskar, B. 2014. Expectation-maximization for learning determinantal point processes. In *NIPS*, 3149–3157.

Gillenwater, J.; Kulesza, A.; and Taskar, B. 2012. Discovering diverse and salient threads in document collections. In *EMNLP-CoNLL*, 710–720.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. In *Neural Computation*, volume 9, 1735–1780.

Huang, L. 2008. Forest reranking: Discriminative parsing with non-local features. In *ACL*, 586–594.

Kang, B. 2013. Fast determinantal point process sampling with application to clustering. In *NIPS*, 2319–2327.

Kulesza, A., and Taskar, B. 2010. Structured determinantal point processes. In *NIPS*, 1171–1179.

Kulesza, A., and Taskar, B. 2011. k-dpps: Fixed-size determinantal point processes. In *ICML*, 1193–1200.

Kulesza, A.; Taskar, B.; et al. 2012. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning* 5(2–3):123–286.

Li, J., and Jurafsky, D. 2016. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, 110–119.

Li, X.; Mou, L.; Yan, R.; and Zhang, M. 2016b. StalemateBreaker: A proactive content-introducing approach to automatic human-computer conversation. In *IJCAI*, 2845–2851.

Mariet, Z., and Sra, S. 2016. Diversity networks: Neural network compression using determinantal point processes. In *ICLR*.

Mou, L.; Song, Y.; Yan, R.; Li, G.; Zhang, L.; and Jin, Z. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *COLING*, 3349–3358.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318.

Ritter, A.; Cherry, C.; and Dolan, W. B. 2011. Data-driven response generation in social media. In *EMNLP*, 583–593.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016a. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, 3776–3783.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A.; and Bengio, Y. 2016b. A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069*.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural responding machine for short-text conversation. In *ACL-IJCNLP*, 1577–1586.

Shao, L.; Gouws, S.; Britz, D.; Goldie, A.; Strope, B.; and Kurzweil, R. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *EMNLP*, 2200–2209.

Song, Y.; Yan, R.; Li, X.; Zhao, D.; and Zhang, M. 2016. Two are better than one: An ensemble of retrieval- and generation-based dialog systems. *arXiv preprint arXiv:1610.07149*.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.-Y.; Gao, J.; and Dolan, B. 2015. A neural network approach to context-sensitive generation of conversational responses. In *NAACL-HLT*, 196–205.

Tian, Z.; Yan, R.; Mou, L.; Song, Y.; Feng, Y.; and Zhao., D. 2017. How to make contexts more useful? an empirical study to context-aware neural conversation models. In *ACL*, 231–236.

Vijayakumar, A. K.; Cogswell, M.; Selvaraju, R. R.; Sun, Q.; Lee, S.; Crandall, D.; and Batra, D. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Vondrák, J.; Chekuri, C.; and Zenklusen, R. 2011. Submodular function maximization via the multilinear relaxation and contention resolution schemes. In *The forty-third annual ACM symposium on Theory of computing*, 783–792. ACM.

Yan, R.; Song, Y.; Zhou, X.; and Wu., H. 2016. "shall i be your chat companion?" towards an online human-computer conversation system. In *CIKM*, 649–658.

Yan, R.; Song, Y.; and Wu., H. 2016. Learning to respond with deep neural networks for retrieval based human-computer conversation system. In *SIGIR*, 55–64.

Yao, J.; Fan, F.; Zhao, W. X.; Wan, X.; Chang, E.; and Xiao, J. 2016. Tweet timeline generation with determinantal point processes. In *AAAI*, 3080–3086.

Yao, L.; Zhang, Y.; Feng, Y.; Zhao, D.; and Yan, R. 2017. Towards implicit content-introducing for generative short-text conversation systems. In *EMNLP*, 2180–2189.

Yin, J.; Jiang, X.; Lu, Z.; Shang, L.; Li, H.; and Li., X. 2016. Neural generative question answering. In *IJCAI*, 2972–2978.

Zhang, M., and Hurley, N. 2008. Avoiding monotony: improving the diversity of recommendation lists. In *The 2008 ACM conference on Recommender systems*, 123–130. ACM.

Zhao, T.; Zhao, R.; and Eskenazi, M. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*, 654–664.