# Modeling Dialogues with Hashcode Representations: A Nonparametric Approach

**Sahil Garg,**[1] **Irina Rish,**[2] **Guillermo Cecchi**[2]

sahil.garg.cs@gmail.com, {rish, gcecchi}@us.ibm.com

**Palash Goyal,**[1] **Sarik Ghazarian,**[1] **Shuyang Gao,**[1] **Greg Ver Steeg,**[1] **Aram Galstyan**[1]

{palashgo, sarikgha, gaos}@usc.edu, {galstyan, gregv}@isi.edu

[1]USC Information Sciences Institute
[2]IBM T. J. Watson Research Center

## Abstract

We propose a novel dialogue modeling framework, the first-ever *nonparametric kernel functions* based approach for dialogue modeling, which learns *hashcodes* as text representations; unlike traditional deep learning models, it handles well relatively *small datasets*, while also scaling to large ones. We also derive a novel *lower bound on mutual information*, used as a model-selection criterion favoring representations with better alignment between the utterances of participants in a *collaborative dialogue* setting, as well as higher predictability of the generated responses. As demonstrated on three real-life datasets, including prominently psychotherapy sessions, the proposed approach significantly outperforms several state-of-art neural network based dialogue systems, both in terms of computational efficiency, reducing training time from days or weeks to hours, and the response quality, achieving an order of magnitude improvement over competitors in frequency of being chosen as the best model by human evaluators.

## 1  Introduction

[1] Dialogue modeling and generation is an active research area of great practical importance as it provides a solid basis for building successful conversational agents in a wide range of applications. However, despite recent successes of deep neural dialogue models, the open dialog generation problem is far from being solved  (Li et al. 2017; Zhao, Lee, and Eskenazi 2018; Wu et al. 2018; Zhang et al. 2018; Pandey et al. 2018; Tao et al. 2018; Xing et al. 2018; Li et al. 2019).

Therefore, it is important to continue exploring novel types of models and model-selection criteria, beyond today's deep neural dialogue systems, in order to better capture the structure of different types of dialogues and to overcome certain limitations of neural models, including dependence on large training datasets, long training times, and difficulties incorporating non-standard objective functions, among others. Along those lines, in this paper, we propose a *first-ever nonparametric approach* (Wasserman 2006), i.e. based on *convolution kernel similarity* functions (Haussler 1999;

Mooney and Bunescu 2005; Scholkopf and Smola 2001), for modeling and generation of dialogues.

Moreover, different applications may possess specific properties which suit some approaches better than others. In this work, one of the motivating applications is a fast-growing area of (semi-)automated *psychotherapy*: easily accessible, round-the-clock psychotherapeutic services provided by a conversational agent. The importance of this area cannot be underestimated: according to recent statistics, mental health disorders affect one in four adult Americans, one in six adolescents, and one in eight children; predicted by the World Health Organization, by 2030 the amount of worldwide disability and life loss attributable to depression may become greater than for any other condition, including cancer, stroke, heart disease, accidents, and war.

However, many people do not receive an adequate treatment. One of the major factors here is limited availability of mental health care professionals, as compared to the number of potential patients; thus, automating at least some aspects of the treatment is a promising direction.

One of the domain-specific challenges in automated therapy is *difficulty obtaining large training datasets* which are often necessary for neural dialogue models; this limitation may require developing alternative approaches. Another domain-specific property of therapeutic dialogues, which can potentially simplify dialogue generation, is the classical pattern of relatively long patient's utterances (up to thousands of words) followed by much shorter therapist's responses. Therapist's responses are often high-level, generic statements, confirming and/or summarizing patient's responses; they *can be viewed as semantic "labels" to be predicted from patient's "input samples"*.

Furthermore, a therapy session is typically an example of a *collaborative dialogue*, unlike debates, political arguments, and so on. Indeed, a fundamental concept in psychotherapy is the working alliance between the therapist and the patient (Bordin 1979). The alliance involves the agreement on the goals to be achieved and the tasks to be carried out, and the bond, trust and respect to be established over the course of the therapy. While an encompassing formalization of working alliance is a challenging task, we propose *maximizing mutual information between the patient's and*

[1]See an extended version of the paper here: arxiv.org/abs/1804. 10188

**Step1**: Learn Hashcode Representations of Dialog Utterances

Maximize Mutual Information

**Patient:** Well I talked to a friend of mine whom I think is promiscuous and she told me about a man that came to see her at her apartment. So I gave him my phone number. I don't have any hang up about morals. I don't know. I told a couple of my friends about this. **I must admit I'm a little shy talking about it now.**

101001101001 ↔ 101010101100

**Therapist**: I see. You mean it's embarrassing to tell me about it.

**Patient:** I'm sitting here just thinking about things. **Should I think out loud?**

100100011101 ↔ 111100001010

**Therapist**: Sure. If you want to.

**Step2**: Infer Hashcode for a Therapist Response using Decision Trees

| Patient Hashcodes (Feature Vectors) | Therapist Hashcodes (Class Labels) |
|---|---|
| 1010001000011 | 101010101100 |
| 1001010010101 | 100100001110 |
| 1001001100101 | 110110100010 |
| 1001100001001 | 010110110010 |
| 1001010011100 | 000100001110 |

**Step3**: Map the Inferred Therapist Hashcode to a Textual Response

100110100101010100100101010
Inferred Therapist Hashcode

Generative Language Model Via N-grams Hashing

Token$_1$  Token$_2$  ...  Token$_N$          Token$_{N+1}$
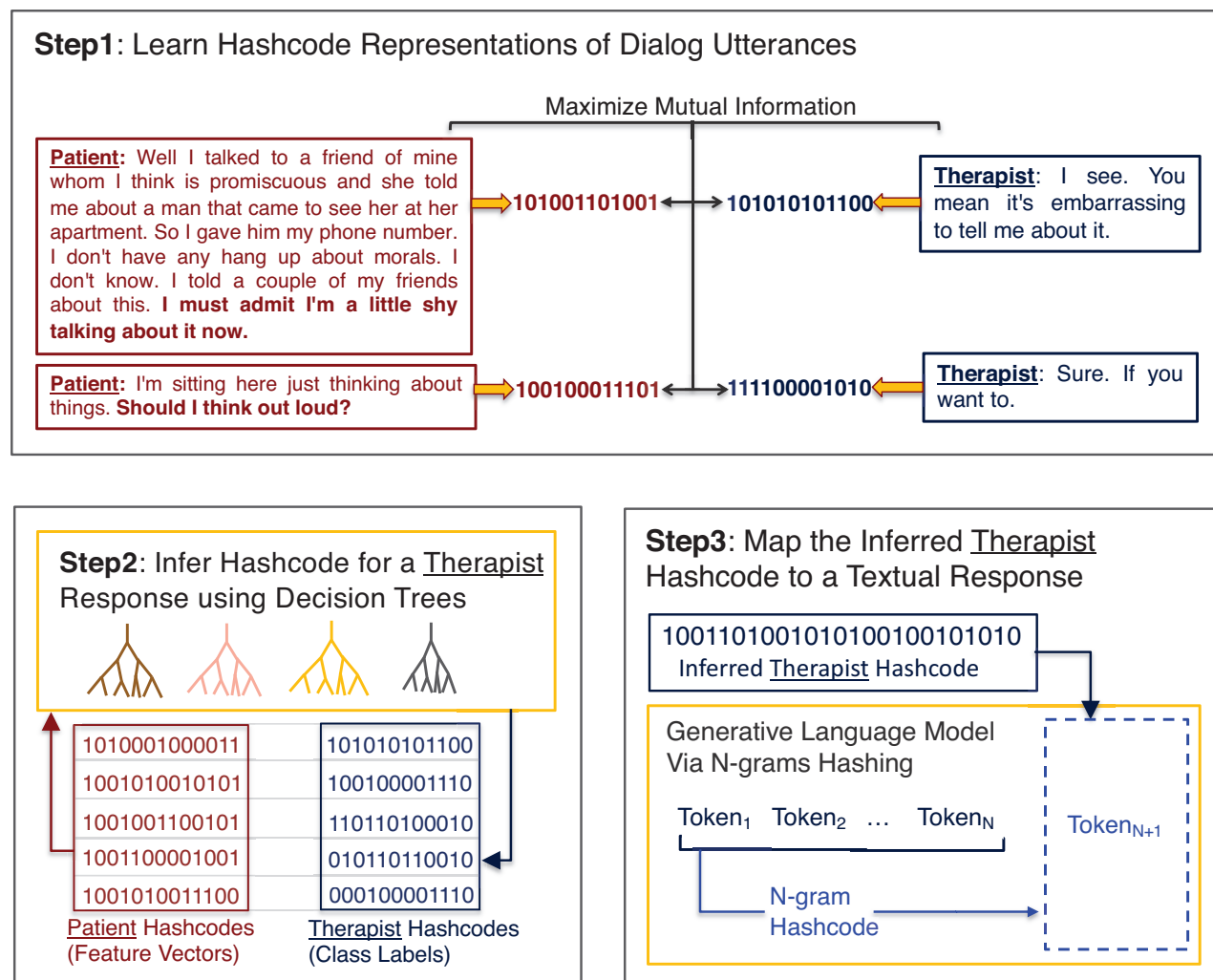
N-gram Hashcode

Figure 1: An intuitive illustration of the kernelized hashcodes based framework for dialog modeling. It involves three steps. In the first step (top), we learn a hashing model to obtain hashcode representations of dialogue utterances; the objective behind learning a hashing representation via maximizing the mutual information between a patient (left) and therapist (right) responses is to find a compressed encoding of those responses which preserves the mutually relevant content while ignoring irrelevant details; e.g., in the example above, we would expect a good representation model to capture the content highlighted in boldface as essential to the conversation. In step 2 (bottom left), we train a ensemble of decision tree classifiers so as to infer hashcode for a therapist response, given the hashcode representation of the corresponding patient utterance (input). In step 3 (bottom right), having the inferred therapist hashcode, we map it to a textual response. One can choose a response from the training set of therapist responses, the one for which its hashcode has low hamming distance w.r.t. the inferred therapist hashcode. Or, we can generate a novel response using N-Grams based language modeling, generalized with kernelized hashing of N-Grams.

*therapist's responses* as a simple criterion helping us capture, to some extent, the dynamics of agreement expected to develop in most therapies, and, more generally, in other types of dialogues[2]. Furthermore, *maximizing mutual information* between the patient's and therapist's can *improve the predictability of the latter from the former*, thus facilitating better dialogue generation.

Motivated by above considerations, we introduce here a novel dialogue modeling framework, the *first-ever nonparametric approach for dialogue modeling*, where responses are represented as *locality-sensitive kernelized hashcodes* (Joly and Buisson 2011; Garg et al. 2019b; 2019a), and the hashing models are optimized using a *novel mutual-information lower bound*, since exact mutual information computation is intractable in high-dimensional spaces; one can also use neural networks for learning the hashcode representations, as we demonstrate in the experiments section. Using hashcode representations allows for a more tractable way of predicting responses in a compressed, general representation space instead of direct generation of textual responses. (Previously, kernelized hashcode representations were successfully applied in the prior works on information-extraction (Garg et al. 2019b; 2019a).) Once the compressed representation of the response is inferred, any separately trained generative model can be plugged in to produce an actual textual response. See Fig. 1 for an illustration of the framework. Note that separating response inference in the representation space from the actual text generation increases method's flexibility, while mutual information criterion facilitates better alignment between the responses of two subjects and higher predictability of the proper response. It is also important to note that, while the psychotherapy domain was our primary motivation, the proposed approach is generally applicable to a wider range of domains as demonstrated in empirical section.

Overall, our *key contributions* include: (1) a novel generic framework for dialogue modeling and generation using *kernelized locality sensitive hash functions*; (2) a *novel lower bound on the Mutual Information (MI)* between the hashcodes of the responses from the two agents used as an optimization criterion for the locality sensitive hash functions; (3) a language model to generate therapist responses, generalizing from the traditional N-grams based language modeling approach via kernelized hashing of N-grams.

We provide an extensive empirical evaluation on three different dialogue domains, from depression therapy to TV show interviews and Twitter data, demonstrating *advantages of our approach* when compared with the state-of-art neural network based dialog systems, both in terms of the *higher quality of generated responses* (especially on relatively small datasets (thousands of samples), including therapy sessions and Larry King TV interviews), as well as *computational efficiency, reducing the training time from days or*

even weeks (e.g. on near-million-sample Twitter dataset) to a few hours.

## 2  Dialog Modeling via Hashcode Representations

We now present a novel framework for dialogue modeling using binary hash functions. We will refer to the two dialogue agents as to a patient and a therapist, respectively, although the approach is generally applicable to a wider variety of dialogue settings, as demonstrated later in the empirical section on datasets such as TV show interviews and Twitter dialogues.

### 2.1  Problem Formulation and Approach Overview

We consider a dialogue dataset consisting of $N$ samples, $\boldsymbol{S}^{pt} = \{S_i^p, S_i^t\}_{i=1}^N$, where each sample is a pair of a patient and a therapist responses, $S_i^p$ and $S_i^t$, respectively; we will also use the following notation: $\bar{\boldsymbol{S}} = \{S_1^p, \cdots, S_N^p, S_1^t, \cdots, S_N^t\}$, $\boldsymbol{S}^p = \{S_1^p, \cdots, S_N^p\}$, $\boldsymbol{S}^t = \{S_1^t, \cdots, S_N^t\}$. Each response is a natural language structure which can be simply a text, or a text with part of speech tags (PoS), or a syntactic/semantic parsing of the text.

Given response $S_i^p$, the dialogue generation task is to produce the response $S_i^t$. We approach this task as a three-stage problem: first, we learn a representation model, based on *locality sensitive hashing*, which maps each text response $S_i$ into some binary hashcode vector $\boldsymbol{c}_i \in \{0, 1\}^H$; second, we train a classifier to infer the therapist's hashcode $\boldsymbol{c}_i^{t*}$ given the patient's hashcode $\boldsymbol{c}_i^p$, so that the inference takes place in the abstract representation space; *hashcode representation aims at capturing, in a compressed form, the semantic essence of the responses while leaving out irrelevant details*; finally, we produce a textual response based on the predicted hashcode representation.

Our objective is to choose a hashcode-based text representation model so that consecutive responses of the dialogue participants are maximally relevant to each other, as measured by the mutual information between the corresponding hashcode representations; from another perspective, this will also make the response of the second's person more predictable given the first person's response.

### 2.2  Background: Kernelized Locality Sensitive Hashing

The main idea behind locality sensitive hashing is that, data points which are similar to each other as per some features, are assigned hashcodes within a short Hamming distance to each other, and vice versa (Grauman and Fergus 2013; Zhao, Lu, and Mei 2014; Wang et al. 2017). Such hashcodes can be used as generalized representations of language structures, e.g. responses of dialogue participants. There are multiple hash functions proven to be locality sensitive (Wang et al. 2017). Kernelized locality-sensitive hashing approaches have also been developed (Kulis and Grauman 2009; Joly and Buisson 2011), which are recently shown to be applicable for learning representations of natural language (Garg

---

[2]Note that imbalanced response length between the two participants, as well as collaborative property, are shared with some other types of dialogues, e.g., TV show interviews such as Larry King dataset analyzed in this paper, where the guest of a show produces long responses, with the host inserting relatively short comments facilitating the interview.

et al. 2019b; 2019a). These techniques rely on a convolution kernel similarity function $K(S_i, S_j; \boldsymbol{\theta})$ defined for any pair of structures $S_i$ and $S_j$ with kernel parameters $\boldsymbol{\theta}$ (Srivastava, Hovy, and Hovy 2013; Mooney and Bunescu 2005; Haussler 1999).

In order to construct hash functions for mapping textual responses to hashcodes, we will first select from a training dataset a random subset of text structures (responses) $\boldsymbol{S}^R \subset \bar{\boldsymbol{S}}$ of size $|\boldsymbol{S}^R| = M$, called a *reference set*. Further, let $h_l(S_i)$, $l = 1, \cdots, H$, denote a set of $H$ binary-valued hash functions, and let $\boldsymbol{h}(S_i)$ denote vector $\{h_l(S_i)\}_{l=1}^H$. The hashcode representation of response $S_i$ will be given as $\boldsymbol{c}_i = \boldsymbol{h}(S_i)$.

To generate a hash function $h_l(S_i)$, for each bit $l$, we first select a random subset $\boldsymbol{S}_l^R \subset \boldsymbol{S}^R$ of the reference set, $|\boldsymbol{S}_l^R| = 2\alpha$. Next, we assign label 0 to $\alpha$ randomly selected elements of $\boldsymbol{S}_l^R$, and label 1 to the remaining $\alpha$ elements of that set, creating an artificial binary-labeled training dataset [3], which can be now fed into any binary classifier to learn a function $h_l(S_i)$. We generate $H$ such random splits of the reference set, and learn the corresponding $H$ binary classifiers, as hash functions. We also tried kernelized k-nearest neighbor classifier (kNN), resulting into hashing approach we refer to as *LSH-RkNN*.

Overall, to obtain a hashcode of a given response $S_i$, we must compute $M$ kernel similarities, $K(S_i, S_j), \forall S_j \in \boldsymbol{S}^R$. For a limited size $(M)$ of the reference set $\boldsymbol{S}^R$, hashcodes can be computed efficiently, with the computational cost linear in $H$; also, note that LSH techniques described above are easily parallelizable.

Finally, our *LSH-RLSTM* model uses LSTM language model for generating hashcodes; no reference set optimization is required here, since LSTM easily handles large training datasets; however, other hyperparameters, including network's architecture, need to be optimized.

## 2.3 Learning Kernelized Hashcode Representations for Dialog Modeling

Given that each specific hashing model described above involves several model-selection choices, our task will be to optimize those choices using the information-theoretic criterion proposed below.

**Optimizing LSH Model Parameters**
As per the discussion of LSH above, an LSH model involves the function $\boldsymbol{lsh}(.; \boldsymbol{\theta}, \boldsymbol{S}^R)$ for mapping text responses to hashcodes: $\boldsymbol{c}_i = \boldsymbol{lsh}(S_i; \boldsymbol{\theta}, \boldsymbol{S}^R)$, where $\boldsymbol{c}_i = \boldsymbol{h}(S_i) = (h_l(S_i))_{l=1}^H$, and where each hash function $h_l(.)$ is built based on a random subset of $\boldsymbol{S}^R$ using either a kernel (kNN, SVM) or a neural network (LSTM) classifier. For the case of kernel-based LSH, $\boldsymbol{\theta}$ are the parameters of a convolution kernel similarity function $K(S_i, S_j)$. For neural hashing (LSH-RLSTM), $\boldsymbol{\theta}$ refers to the neural architecture hyperparameters (number of layers, the number of units in a layer,

---

[3]For a more sophisticated methodology to assign the artificial binary labels, and to select the random subsets of $\boldsymbol{S}^R$, refer to the recent work by (Garg et al. 2019a).

type of units, etc.); $\boldsymbol{\theta}$ also includes LSH-specific parameters such as $\alpha$.

When learning LSH models on a training datset, the (hyper) parameters $\boldsymbol{\theta}$ as well as the reference set $\boldsymbol{S}^R$ will be optimized with respect to the information-theoretic objective introduced below. Namely, for LSH-RkNN and LSH-RMM, the kernel parameters $\boldsymbol{\theta}$ are optimized via grid search. For LSH-RLSTM, $\boldsymbol{\theta}$ reflects the neural architecture, i.e. the number of layers and the number of units in each layer, optimized by greedy search. Similarly, $\boldsymbol{S}^R$ is also constructed via a greedy algorithm.

**Information-Theoretic Objective Function**
The objective function for hashcode-based model selection in dialog generation should (1) characterize the quality of hashcodes as generalized/compressed representations of dialogue responses and (2) favor representation models leading to higher-accuracy response generation.

Mutual information $\mathcal{I}(S^p : S^t)$ between the dialog responses $S^p$ (e.g. patient) and $S^t$ (e.g., therapist) is a natural candidate objective as it implies higher predictability of one response from another. Though, it is hard to compute in practice as the joint distribution over all pairs of textual responses is not available. However, we can attempt to approximate it using hashcode representations. If $\boldsymbol{h}(.)$ represents a function from the space of all statements to the hashing code space, then the *data processing inequality* implies that $\forall \boldsymbol{h}(.), \mathcal{I}(S^p : S^t) \geq \mathcal{I}(\boldsymbol{h}(S^p) : \boldsymbol{h}(S^t))$, and maximizing the quantity on the right can be more computationally feasible.

Thus we will maximize the *mutual information (MI)* between the response hashcodes, over LSH model parameters; it turns out that MI reflects both the inference accuracy as well as the representation quality, as we will see below:

$$\underset{\boldsymbol{\theta}, \boldsymbol{S}^R}{\arg\max} \ \ \mathcal{I}(\boldsymbol{C^p} : \boldsymbol{C^t}); \tag{1}$$

$$\boldsymbol{C^p} = \boldsymbol{lsh}(S^p; \boldsymbol{\theta}, \boldsymbol{S}^R), \ \boldsymbol{C^t} = \boldsymbol{lsh}(S^t; \boldsymbol{\theta}, \boldsymbol{S}^R) \tag{2}$$

$$\mathcal{I}(\boldsymbol{C^p} : \boldsymbol{C^t}) = \mathcal{H}(\boldsymbol{C^t}) - \mathcal{H}(\boldsymbol{C^t} | \boldsymbol{C^p}) \tag{3}$$

Herein, $\boldsymbol{C^p}$ and $\boldsymbol{C^t}$ are the multivariate binary random variables associated with the hashcodes of patient and therapist responses, respectively. Minimizing the conditional entropy, $\mathcal{H}(\boldsymbol{C^t} | \boldsymbol{C^p})$, improves the predictive accuracy when inferring therapist response hashcode, while maximizing the entropy term, $\mathcal{H}(\boldsymbol{C^t})$, should ensure good quality of the hashcodes as generalized representations of text responses; thus MI objective satisfies both criteria stated at the beginning of this section.

**Mutual Information Lower Bound for Efficiency**
Since computing mutual information between two high-dimensional variables can be both computationally expensive and inaccurate if the number of samples is small (Kraskov and Grassberger 2004; Walters-Williams and Li 2009; Singh and Póczos 2014; Gao, Ver Steeg, and Galstyan 2015), we develop a (novel) lower bound on the mutual information which is easy to compute. For derivation details, see the supplementary material.

**Algorithm 1** Response Generation via Hashing of N-grams

---

**Require:** Therapist response sentences from the training set, $\boldsymbol{S^t}$ and their hashcodes, $\boldsymbol{C^t}$; N-grams from the training set, $\boldsymbol{S^g}$, and their hashcodes, $\boldsymbol{C^g}$; an inferred therapist hashcode, $\boldsymbol{c^t_*}$.

1: $S^t \leftarrow \text{samplePrefix}(\boldsymbol{c^t_*}, \boldsymbol{C^t}, \boldsymbol{S^t})$     *% sampling from therapist responses matching the inferred therapist hashcode.*
2: **while** True **do**
3:    $S^g \leftarrow \text{extractSuffixAsNgram}(S^t)$
4:    $\boldsymbol{c^g} \leftarrow \text{computeHashcodeOfNgram}(S^g)$
5:    $w^g \leftarrow \text{inferNextToken}(\boldsymbol{c^g}, \boldsymbol{C^g})$    *% for an Ngram, there are tokens to follow as suffix along with the probabilities. Tokens from all the relevant N-Grams in the training set, matched via hashcodes, are aggregated for sampling.*
6:    $S^t \leftarrow \text{appendTokenToResponse}(S^t, w^g)$
7:    $\boldsymbol{c^t} \leftarrow \text{computeHashcodeOfResponse}(S^t)$
8:    $f^t \leftarrow \text{computeLikelihood}(\boldsymbol{c^t}, \boldsymbol{c^t_*}, \boldsymbol{C^t})$    *% likelihood per the responses in training set matched via hashcodes, and hamming distance w.r.t. the inferred hashcode.*
9:    **if** $f^t$ is high **then**
10:     **return** $S^t, f^t$    *% return the therapist response.*
11:    **else if** $f^g$ is very low **then**
12:     $S^t \leftarrow \text{backtrack}(S^t)$    *% Some of the tokens in the suffix are removed from the response.*
13:    **end if**
14: **end while**

---

We will first introduce the information-theoretic quantity called *Total Correlation* (Watanabe 1960), $\mathcal{TC}(\boldsymbol{C}) = \sum_j \mathcal{H}(C_j) - \mathcal{H}(\boldsymbol{C})$, which captures non-linear correlation among the dimensions of a random variable $\boldsymbol{C}$; given an additional random variable $\boldsymbol{Y}$, $\mathcal{TC}(\boldsymbol{C} : \boldsymbol{Y})$ is defined as:

$$\mathcal{TC}(\boldsymbol{C} : \boldsymbol{Y}) = \mathcal{TC}(\boldsymbol{C}) - \mathcal{TC}(\boldsymbol{C}|\boldsymbol{Y}). \quad (4)$$

**Theorem 1** (Lower Bound on Mutual Information). *Mutual information between two random hashcode variables, $\mathcal{I}(\boldsymbol{C^p} : \boldsymbol{C^t})$, can be bounded from below as follows:*

$$\mathcal{I}(\boldsymbol{C^p} : \boldsymbol{C^t}) \geq \sum_j \mathcal{H}(C^t_j) - \mathcal{TC}(\boldsymbol{C^t} : \boldsymbol{Y^*}) + \sum_j \big\langle \log q(C^t_j | \boldsymbol{C^p}) \big\rangle.$$

*Herein, $\mathcal{TC}(\boldsymbol{C^t} : \boldsymbol{Y})$ describes Total Correlations within $\boldsymbol{C^t}$ that can be explained by a latent variables representation $\boldsymbol{Y}$; $q(C^t_j | \boldsymbol{C^p})$ is a proposal conditional distribution for the $j_{th}$ bit of the hashcode $\boldsymbol{C^t}$ predicted using a probabilistic classifier, like a Random Forest model.*

As discussed in (Ver Steeg and Galstyan 2014), $\mathcal{TC}(\boldsymbol{C^t} : \boldsymbol{Y^*})$ can be computed efficiently.

Note that the first two terms in the MI lower bound contribute to improving the quality of hashcodes as response representations, maximizing entropy of each hashcode bit while discouraging redundancies between the bits, while the last term containing conditional entropies aims at improving inference of individual hashcode bits.

Moreover, one can use the proposed MI LB as an *evaluation metric of the dialog quality on test data*, i.e. the alignment/relevance between the responses of two dialog agents.

Also, note that mutual information criterion has been used previously to optimize kernelized hashcode representations for binary classification problems (Garg et al. 2019b;

2019a). The approximation of mutual information proposed in those prior works are not applicable for dialog modeling.

We also use a normalized metric, dividing MI LB by an upper bound on joint entropy,

$$\mathcal{H}(\boldsymbol{C^t}) = \sum_j \mathcal{H}(C^t_j) - \mathcal{TC}(\boldsymbol{C^t}) \quad (5)$$

$$\leq \sum_j \mathcal{H}(C^t_j) - \mathcal{TC}(\boldsymbol{C^t} : \boldsymbol{Y^*}). \quad (6)$$

For $\mathcal{TC}(\boldsymbol{C^t}|\boldsymbol{Y^*}) = 0$, i.e. when a latent representation $\boldsymbol{Y^*}$ is learned which explains all the Total Correlations in $\boldsymbol{C_t}$, the upper bound becomes equal to the entropy term; practically, for the case of hashcodes, learning such a representation should not be difficult, so the bound should be tight.

Having learned a model to hash patient or therapist utterances, we train a Random Forest classifier, or any classifier, to infer hashcode for a therapist response given the input of a hashcode of a patient utterance, as illustrated in Fig. 1. For mapping the inferred therapist hashcode to a textual response, we have two choices. One simple choice, which works well for therapeutic dialog like problems, is to select a therapist response from the training set by matching its hashcode w.r.t. the inferred hashcode. Second choice is to generate a new response given the inferred hashcode, for which we propose a novel approach in the following section.

## 2.4 Dialogue Generation with Kernelized Hashing of N-grams

For generation of a therapist dialogue response, given the input of its kernelized hashcode representation, we propose a new language model by generalizing the traditional approach of N-grams language modeling with kernelized hashing of N-grams. The fundamental problem with the tradi-

| Model | Average | Greedy | Extrema | Average | Greedy | Extrema | Average | Greedy | Extrema |
|---|---|---|---|---|---|---|---|---|---|
| LSTM (Vinyals and Le 2015) | 0.61±0.31 | **0.58±0.29** | 0.28±0.16 | 0.51 | 0.39 | 0.37 | 0.71±0.24 | 0.60±0.20 | 0.35±0.14 |
| HRED(Serban et al. 2016a) | 0.48±0.23 | 0.43±0.20 | 0.29±0.16 | 0.50 | 0.38 | 0.36 | 0.71±0.25 | 0.61±0.20 | 0.29±0.12 |
| VHRED(Serban et al. 2017) | 0.48±0.23 | 0.43±0.20 | 0.29±0.16 | 0.53 | 0.40 | **0.38** | 0.70±0.24 | **0.72±0.25** | **0.43±0.18** |
| LSH-RkNN | 0.60±0.37 | **0.58±0.34** | **0.37±0.24** | **0.61±0.17** | 0.40±0.13 | 0.25±0.09 | **0.76±0.28** | 0.60±0.21 | 0.34±0.15 |
| LSH-RMM | 0.56±0.38 | 0.53±0.33 | 0.31±0.23 | **0.61±0.17** | **0.41±0.13** | 0.25±0.09 | 0.73±0.28 | 0.59±0.22 | 0.35±0.16 |
| LSH-RLSTM | **0.64±0.37** | 0.51±0.28 | 0.28±0.19 | 0.60±0.18 | 0.39±0.13 | 0.24±0.09 | **0.76±0.27** | 0.58±0.21 | 0.33±0.15 |

| (a) Depression Therapy Dataset | (b) Twitter Dataset | (c) Larry King Dataset |
|---|---|---|

Table 1: Comparison between state-of-art neural network models (LSTM, HRED and VHRED) and the proposed hashing models (LSH-RkNN, LSH-RMM and LSH-RLSTM), on three datasets – Depression Therapy, Twitter, and Larry King data – using word embedding-based similarity metrics between the actual and generated responses. Mean and standard deviation across samples (response pairs) are reported for all metrics, for each test set except for Twitter results with prior art models (LSTM, HRED, VHRED) - we used the numbers reported in (Serban et al. 2017), without rerunning the models; standard deviations were not reported in that paper.

tional N-grams based language models is that the approach is restricted to N-grams of short length (Pauls and Klein 2011; Bengio et al. 2003). As we increase the length of N-grams, it becomes difficult to generalize the applicability of N-grams from a training corpus to test settings due an explosion in the number of possible N-grams.

In light of the recent developments on learning kernelized hashcode representations of natural language (Garg et al. 2019b; 2019a), it is possible to generalize the existing N-gram language models via kernelized hashing of N-grams. While it is low probable to exactly match an N-gram of large length w.r.t. the ones in a training corpus, it is feasible to find a match for its hashcode representation. This is because N-grams sharing relevant patterns are assigned same kernelized hashcode. Along these lines, instead of N-grams, probabilities are computed for hashcodes of N-grams per the occurrence statistics of N-grams assigned same hashcode, and so it applies for the conditional probabilities, which gives us the basic language model for generating therapist responses; see the pseudo code in Alg. 1.

For hashing of N-grams, we use an unsupervised kernelized hashing model; the supervised dialog hashing model, that we proposed in the previous section, is used only for hashing of dialogue utterances, not N-grams. In (Garg et al. 2019a), an information theoretic approach for nearly-unsupervised learning of kernelized representations is proposed for information extraction task, that is easily extensible for unsupervised learning settings.

## 3 Empirical Evaluation

Several variants of the proposed hashing based dialog model, using kNN, SVM or LSTM to build hashcodes, respectively, were evaluated on three different datasets and compared with three state-of-art dialog generation approaches of (Serban et al. 2017; 2016a) and (Vinyals and Le 2015). Besides several standard evaluation metrics adopted by those approaches, we also report the *model rankings obtained by human evaluators* via Amazon Mechanical Turk.

### 3.1 Experimental Setup

**Datasets**

The three datasets used in our experiments include (1) depression therapy sessions, (2) Larry King TV interviews and (3) Twitter dataset. The *depression therapy dataset*[4] consists of transcribed recordings of nearly 400 therapy sessions between multiple therapists and patients. Each patient response $S_i^t$ followed by therapist response $S_i^p$ is treated as a single sample; all such pairs, from all sessions, were combined into one set of N=42000 samples. We select 10% of the data randomly as a test set (4200 samples), and then perform another random 90/10 split of the remaining 38,000 samples into training and validation subsets, respectively. We follow the experimental setup from prior work cited above when comparing the respective neural network models with our hashing based approaches: all models are trained only once using the same training and validation datasets, and evaluated on the same test set. However, for our hashing model metrics introduced below, we average the estimates over 10 random subsets using 95% of test samples each time.

The *Larry King dataset* [5] contains transcripts of interviews with the guests of TV talk shows, conducted by Larry King, the host. Similarly to the depression therapy dataset, we put together all pairs of guest/host responses from 69 sessions into a single set of size 8200. The data are split into training, validation and test subsets as described earlier.

Next, we experimented with the *Twitter Dialogue Corpus* (Ritter, Cherry, and Dolan 2010). Considering the original tweet and the following comments on it, in the same session, the task is to infer the next tweet. Note that we consider all utterances preceding that tweet as one long utterance, i.e. as the first "response" $S_i^A$, mapped to one hashcode, while the next tweet is the second "response" $S_i^B$, which is different from the approach of (Serban et al. 2017) we compare with, where the previous utterances in a session are explicitly viewed as a sequence. The number of tweet ses-

---

[4]https://alexanderstreet.com/products/counseling-and-psychotherapy-transcripts-series

[5]http://transcripts.cnn.com/TRANSCRIPTS/lkl.html

|  | LSTM | HRED | VHRED | LSH-RkNN | LSH-RMM | LSH-RLSTM |
|---|---|---|---|---|---|---|
| Appropriate (%) | 3.7 | 8.6 | 9.5 | **28.7** | 24.1 | 25.4 |
| Diverse (%) | 0.7 | 9.3 | 0.7 | 35.1 | 13.2 | **41.1** |

(a) Depression Dataset

|  | LSTM | HRED | VHRED | LSH-RkNN | LSH-RMM | LSH-RLSTM |
|---|---|---|---|---|---|---|
| Appropriate (%) | 5.3 | 3.9 | 5.3 | **31.7** | 25.6 | 28.3 |
| Diverse (%) | 3.3 | 13.3 | 0.0 | **36.7** | 10.0 | **36.7** |

(b) Larry King Dataset

Table 2: Human evaluation scores on (a) Depression dataset (900 test samples) and (b) Larry King dataset (180 test samples).

| Model | MI LB (Shuffled) | NMI LB | HIA (Baseline) |
|---|---|---|---|
| LSH-RkNN | 12.8±0.5 (6.1±0.2) | 0.57 (0.27) | 0.87±0.16 (0.82±0.18) |
| LSH-RMM | **13.7±0.2 (-1.0±0.3)** | **0.39 (0.0)** | **0.68±0.10 (0.59±0.10)** |
| LSH-RLSTM | 20.3±0.3 (10.6±0.2) | 0.76 (0.40) | 0.82±0.19 (0.79±0.21) |

(a) Depression Therapy Dataset

| Model | MI LB (Shuffled) | NMI LB | HIA (Baseline) |
|---|---|---|---|
| LSH-RkNN | 17.3±0.1 (9.1±0.1) | 0.75 (0.40) | 0.85±0.18 (0.81±0.21) |
| LSH-RMM | 23.4±0.2 (11.2±0.3) | **0.71 (0.34)** | **0.80±0.14 (0.72±0.16)** |
| LSH-RLSTM | **41.9±0.1 (24.3±0.1)** | 0.83 (0.48) | 0.69±0.16 (0.63±0.16) |

(b) Twitter Dataset

| Model | MI LB (Shuffled) | NMI LB | HIA (Baseline) |
|---|---|---|---|
| LSH-RkNN | 9.4±0.3 (5.1±0.3) | 0.63 (0.34) | 0.91±0.14 (0.89±0.16) |
| LSH-RMM | 22.4±0.9 (4.5±1.2) | **0.62 (0.13)** | **0.69±0.11 (0.59±0.10)** |
| LSH-RLSTM | **48.9±0.3 (28.4±0.5)** | 0.80 (0.47) | 0.62±0.10 (0.54±0.07) |

(c) Larry King Dataset

Table 3: Hashcode models quality as measured by the alignment between the hashcodes of person A and person B responses (mutual information lower bound and its normalized version), as well as by the predictability of B's responses given A's response.

sions (each viewed as a separate sample, i.e. $\{S_i^A, S_i^B\}$ pair of responses), in training, validation, and test subsets are, respectively, 749060, 93633, and 93633.

**Task**
For all datasets, the task is to train a model on a set of training samples, i.e. response pairs $(S_i^A, S_i^B)$, where $S_i^A$ is a response of person A, followed by the corresponding response of a person B. Then each test sample is given as a response of person A, and the task is to generate the response of a person B.

**Hashing Models**
*Step 1: Representation Learning.* We evaluate three different hashing models: the first two, based on kernel locality sensitive hashing (KLSH) (Joly and Buisson 2011; Garg et al. 2019b), are called *LSH-RMM* and *LSH-RkNN*, and use, respectively, Max-Margin (SVM) classifier (with C=1 parameter) or kNN classifier (k=1), to compute each hash function. The third model, *LSH-RLSTM*, uses LSTM for hash function computation. We use hashcode vectors of dimensionality H=100. For *LSH-RkNN* and *LSH-RMM*, we use as a reference set a random subset of M=300 samples from the training dataset, to reduce the computational complexity of training those models, but for *LSH-RLSTM* we use the whole training dataset as a reference set. Parameters $\theta$ for LSH models are obtained by maximizing the proposed MI LB criterion.

*Step 2: Hashcode prediction.* We now map all responses, of both participants A and B, in both training and test sets, to the corresponding hashcodes using one of the above hashcode-based representation models. Next, to predict the response hashcode of a person B given a hashcode of a person A, we train separate Random Forest (RF) classifiers (each containing 100 decision trees) for each hashcode bit (i.e. 100 such RF classifiers, since H=100).

*Step 3: Textual response generation.* Given a hashcode of a response inferred by RF classifier above, mapping it to an actual text can be performed in multiple ways; by default, we find a match of the generated hashcode in the set of all hashcodes corresponding to the person B responses in our training data as this approach is highly suitable for depression therapy like dialogues. In addition, we also provide human evaluation for generating new responses using the N-grams hashing based language model proposed in Sec. 2.4.

**Baseline: Neural Network Dialog Generation Models**
We compare our dialog generation method with the state-of-art *VHRED* approach of (Serban et al. 2017), as well as with the two other approaches, *HRED* (Serban et al. 2016a), and *LSTM* (Vinyals and Le 2015), also used as baselines in the VHRED paper. We adopt the same hyperparameter settings as those used in (Serban et al. 2017). For the Twitter dataset, we compare with the results presented in the above paper, while on the other two datasets, we train the above models ourselves. The vocabulary size for the input is set via grid search between values 1000 to 100000. The neural network structures are chosen by an informal search over a set of architectures and we set maximum gradient steps to 80, validation frequency to 500 and step-size decay for SGD is 1e-4.

**Evaluation Metrics**
*Embedding-based metrics.* We compare our methods with the state-of-art neural network approaches listed above using three word embedding-based topic similarity metrics - *embedding average*, *embedding greedy*, and *embedding extrema* (Liu et al. 2016), adopted by (Serban et al. 2017). Following the prior art, we used Google News Corpus to train the embeddings. The mean and standard deviation statistics for each metric are computed over 10 runs of the experiment, as mentioned above.

*Human evaluation.* Using Amazon Mechanical Turk, we obtained model rankings from 108 human readers (annotators). For each test sample, we showed to the reader all responses produced the six models evaluated here, in random order for each instance, and without specifying which model produced which response. We asked the annotator to choose two most appropriate responses; then, for each model, we com-

|  | LSTM | HRED | VHRED | LSH-RkNN | **LSH-RkNN-Gen** |
|---|---|---|---|---|---|
| Appropriate (%) | 3.7 | 12.2 | 13.5 | 36.8 | 33.8 |
| Diverse (%) | 0.0 | 4.0 | 4.0 | 80 | 12.0 |

Table 4: Human evaluation scores on Depression dataset (300 test samples) for our generative hashing model (LSH-RkNN-Gen) w.r.t. the baseline models, and our response selection based hashing model (LSH-RkNN).

puted across all test samples the percent of the readers who voted for that model. Furthermore, in a separate session, we also asked the person which model produced the most diverse responses; in this case we listed the model's names and associated responses.

*Information-theoretic metrics.* The second group of metrics directly evaluates the quality of hashcodes obtained using our models. For each method, we will report the proposed MI lower bound (*MI LB*) from Theorem 1, as well as its normalized version (NMI LB). For each of the metrics, *higher values* mean better performance. We also report the *hashcode inference accuracy (HIA)*, i.e. the accuracy of predicting hashcode bits of response B using RF classifiers. We also obtain the baseline accuracy (*Baseline*), using a trivial classifier that always chooses the most-frequent class label.

## 3.2 Results

### Computational Efficiency
First of all, we observed that *the hashing models are much more computationally efficient than the neural network approaches we compared with*: it takes from several days (for smaller datasets such as Depression therapy and Larry King dataset containing about 42,000 and 8,200 samples) to more than two weeks (on Twitter, with roughly 749,000 samples) to train a neural network model, even on a 1000-core GPU, whereas an LSH model is typically optimized within several hours (e.g., less than 5-6 hours for Twitter), on a 16-core CPU.

### Embedding-based Metrics
Table 1 summarizes performance of all methods with respect to the embedding-based metrics, with the best results for each metric/column shown in boldface. Overall, the proposed hashing approaches are quite competitive with the neural net methods in terms of those metrics as well: on all three datasets, our LSH methods always outperform their competitors in terms of the *average* similarity metric (which is the most intuitive among the three metrics in terms of reflecting the topic similarity between the true and system-generated responses); moreover, our methods also achieve best performance w.r.t. the *greedy* metric on Twitter dataset, and w.r.t. *extrema* metric on the Depression dataset.

### Human Evaluation
We also performed an extensive human evaluation of the dialog generation quality (as described earlier in section 5.1), on two relatively smaller datasets, Depression and Larry King. The results are shown in Table 2a and 2b, respectively. *The hashcode-based approaches considerably outperform the neural net models, by one-two orders of mag-*

*nitude, in terms of the responses being both more appropriate and more diverse*[6]. Amongst the three hashing models, LSH-RkNN seem to be performing the best.

Further, we performed another human evaluation of LSH-RkNN, for a comparison between response generation vs response selection, and their comparison w.r.t. the baseline neural models. As we observe in Tab. 4, both response selection and response generation seem to give equally good responses in terms of appropriateness, and significantly outperform the neural baselines. Low diversity of generated responses remains an issue, as also known to be the case for neural models.

### Information-theoretic Metrics
Next, we took a deep dive into evaluation of hashing approaches with respect to how well they actually model the alignment between the responses; the results are summarized in Table 3, presenting the mutual information lower bound (MI LB), normalized MI LB, and hashcode inference accuracy (HIA) as discussed before; note that besides presenting chance-level baseline classifier accuracy (i.e., selecting the most-frequent class), we also present a similar baseline for MI LB and NMI LB metrics, using randomly permuted (shuffled) pairings between responses of person A and person B. We now look at the difference between each metric and its corresponding baseline, shown in parenthesis, and highlight the largest, most significant differences between the means of both metrics in boldface.

Overall, *LSH-RMM* method appear to be the best on all datesets in terms of most significant improvement over the baseline in terms of normalized MI LB and prediction accuracy, although *LSH-RLSTM* is the best in term of the raw MI LB on Twitter and Larry King datasets.

## 4 Related Work
Therapy chatbots, such as Woebot (Fitzpatrick, Darcy, and Vierhile 2017) and similar systems, are becoming increasingly popular; however, these agents have limited ability to understand free text and have to resort to a fixed set of pre-programmed responses to choose from (Di Prospero et al. 2017; Ly, Ly, and Andersson 2017; Schroeder et al. 2018; Morris et al. 2018; Hamamura et al. 2018).

For dialogue modeling in general domains, several recently proposed neural network based approaches are considered state-of-art (Serban et al. 2016b; 2017; Zhao, Lee, and Eskenazi 2018; Tao et al. 2018; Li et al. 2019). However, those approaches usually require very large training datasets, unavailable in many practical applications; furthermore, they are not typically explored in dialogue settings such as therapy including very long responses (up to tens of thousands of words). Also, evaluating the effectiveness of the therapist's response requires some notion of relevance (e.g., mutual information) which goes beyond the standard measures of its semantic features (Papineni et al. 2002; Liu et al. 2016; Li and Jurafsky 2017; Lowe et al. 2017).

Unlike task-driven dialogue (Zhai and Williams 2014; Wen et al. 2017; Althoff, Clark, and Leskovec 2016; Lewis

---

[6]See several examples in the extended version.

et al. 2017; He et al. 2018), an immediate response quality metric may not be available in our settings, since the effect of therapy is harder to evaluate and multiple sessions are often required to achieve the desired outcome. Attention to specific parts of the response, as well as background knowledge, explored in neural network based dialogue modeling (Kosovan, Lehmann, and Fischer 2017) can be helpful in therapeutic dialogues; those aspects are, to some extent, implicitly captured by learning the hashing models.

In (He et al. 2018), an approach is presented to a task-driven (e.g., negotiation) dialogue which consists in mapping a response to an ordered list of rules, where each rule represents a task-specific intent; this does not apply to our more open-ended dialogues without the specific tasks.

Finally, there are multiple approaches for estimating mutual information from data (Kraskov and Grassberger 2004; Koeman and Heskes 2014). However, these estimators are highly expensive in high-dimensional settings, and can be quite inaccurate when the number of samples is small. There is a recent approach of neural estimation of mutual information on two high dimensional continuous variables, though not applicable to discrete variables like hashcodes (Belghazi et al. 2018). For discrete variables, theoretical analysis has been limited to one dimensional case (Jiao et al. 2017).

## 5   Conclusions

This paper introduces a novel approach to dialogue modeling where responses of both participants are represented by kernelized hashcodes. Furthermore, a novel lower bound on Mutual Information is derived and used as a hascode-based model-selection criterion in order to facilitate a better alignment in collaborative dialogue, as well as predictability of responses. Our empirical results consistently demonstrate superior performance of the proposed approach over state-of-art neural network dialogue models in terms of both computational efficiency and response quality. For next steps, we plan to further improve the approach by choosing a better response from a larger corpus, implement more sophisticated hashcode-to-text generative models, and extend the model to the dynamics of a dialogue beyond immediate responses. Additionally, we will exploit the vast literature on operationalized assessment of several psychological qualities such as the Beck Depression (Beck AT 1961) and Beck Anxiety (Beck AT 1988) Inventories, as well as features of dialogue such as the Working Alliance Inventory (Horvath 1986), in order to derive better hashcode representations and predictive models. In this sense, we expect that the lessons learned in richly evaluated therapy sessions can be extrapolated, mutatis mutandis, beyond the realm of mental health.

## References

Althoff, T.; Clark, K.; and Leskovec, J. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *TACL*.

Beck AT, Ward CH, M. M. M. J. E. J. 1961. An inventory for measuring depression. *Arch. Gen. Psychiatry*.

Beck AT, Epstein N, B. G. S. R. 1988. An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology*.

Belghazi, M. I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, R. D. 2018. Mine: mutual information neural estimation. *In Proc. of ICML*.

Bengio, Y.; Ducharme, R.; Vincent, P.; and Jauvin, C. 2003. A neural probabilistic language model. *Journal of machine learning research*.

Bordin, E. S. 1979. The generalizability of the psycho-analytic concept of the working alliance. *Psychotherapy: Theory, research & practice*.

Di Prospero, A.; Norouzi, N.; Fokaefs, M.; and Litoiu, M. 2017. Chatbots as assistants: an architectural framework. In *Proc of the 27th Annual Intl Conf on Computer Science and Software Engineering*. IBM Corp.

Fitzpatrick, K.; Darcy, A.; and Vierhile, M. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*.

Gao, S.; Ver Steeg, G.; and Galstyan, A. 2015. Efficient estimation of mutual information for strongly dependent variables. In *AISTATS*.

Garg, S.; Galstyan, A.; Ver Steeg, G.; and Cecchi, G. 2019a. Nearly-unsupervised hashcode representations for relation extraction. In *Proc. of EMNLP*.

Garg, S.; Galstyan, A.; Ver Steeg, G.; Rish, I.; Cecchi, G.; and Gao, S. 2019b. Kernelized hashcode representations for relation extraction. In *Proc. of AAAI*.

Grauman, K., and Fergus, R. 2013. Learning binary hash codes for large-scale image search. *Machine learning for computer vision*.

Hamamura, T.; Suganuma, S.; Ueda, M.; Mearns, J.; and Shimoyama, H. 2018. Standalone effects of a cognitive behavioral intervention using a mobile phone app on psychological distress and alcohol consumption among japanese workers: Pilot nonrandomized controlled trial. *JMIR Mental Health*.

Haussler, D. 1999. Convolution kernels on discrete structures. Technical report.

He, H.; Chen, D.; Balakrishnan, A.; and Liang, P. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proc. of EMNLP*.

Horvath, A. O., G. L. 1986. *The development of the Working Alliance Inventory: A research handbook*.

Jiao, J.; Venkat, K.; Han, Y.; and Weissman, T. 2017. Maximum likelihood estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*.

Joly, A., and Buisson, O. 2011. Random maximum margin hashing. In *CVPR*.

Koeman, M., and Heskes, T. 2014. Mutual information estimation with random forests. In *ICONIP*.

Kosovan, S.; Lehmann, J.; and Fischer, A. 2017. Dialogue response generation using neural networks with attention

and background knowledge. In *Proceedings of the Computer Science Conference for University of Bonn Students (CSCUBS)*.

Kraskov, A.and Stögbauer, H., and Grassberger, P. 2004. Estimating mutual information. *Physical Review E*.

Kulis, B., and Grauman, K. 2009. Kernelized locality-sensitive hashing for scalable image search. In *ICCV*.

Lewis, M.; Yarats, D.; Dauphin, Y.; Parikh, D.; and Batra, D. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proc. of EMNLP*.

Li, J., and Jurafsky, D. 2017. Neural net models of open-domain discourse coherence. In *Proc. EMNLP*.

Li, J.; Monroe, W.; Shi, T.; Jean, S.; Ritter, A.; and Jurafsky, D. 2017. Adversarial learning for neural dialogue generation. In *Proc. of EMNLP*.

Li, J.; Qiu, L.; Tang, B.; Chen, D.; Zhao, D.; and Yan, R. 2019. Insufficient data can also rock! learning to converse using smaller data with augmentation. In *Proc. of AAAI*.

Liu, C.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *In Proc. of EMNLP*.

Lowe, R.; Noseworthy, M.; Serban, I.; Angelard-Gontier, N.; Bengio, Y.; and Pineau, J. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. *In Proc. of ACL*.

Ly, K.; Ly, A.; and Andersson, G. 2017. A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. *Internet Interventions*.

Mooney, R., and Bunescu, R. 2005. Subsequence kernels for relation extraction. In *Proc. of NIPS*.

Morris, R.; Kouddous, K.; Kshirsagar, R.; and Schueller, S. 2018. Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *Journal of Medical Internet Research*.

Pandey, G.; Contractor, D.; Kumar, V.; and Joshi, S. 2018. Exemplar encoder-decoder for neural conversation generation. In *Proc. of ACL*.

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

Pauls, A., and Klein, D. 2011. Faster and smaller n-gram language models. In *Proc. of ACL*.

Ritter, A.; Cherry, C.; and Dolan, B. 2010. Unsupervised modeling of Twitter conversations. In *Proc. of ACL*.

Scholkopf, B., and Smola, A. J. 2001. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Schroeder, J.; Wilkes, C.; Rowan, K.; Toledo, A.; Paradiso, A.; Czerwinski, M.; Mark, G.; and Linehan, M. 2018. Pocket skills: A conversational mobile web app to support dialectical behavioral therapy. In *CHI*.

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

Serban, I.; Lowe, R.; Charlin, L.; and Pineau, J. 2016b. Generative deep neural networks for dialogue: A short review. *arXiv preprint arXiv:1611.06216*.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A. C.; and Bengio, Y. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proc. of AAAI*.

Singh, S., and Póczos, B. 2014. Generalized exponential concentration inequality for Rényi divergence estimation. In *Proc. of ICML*.

Srivastava, S.; Hovy, D.; and Hovy, E. 2013. A walk-based semantically enriched tree kernel over distributed word representations. In *Prof. of EMNLP*.

Tao, C.; Gao, S.; Shang, M.; Wu, W.; Zhao, D.; and Yan, R. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *Proc. of IJCAI*.

Ver Steeg, G., and Galstyan, A. 2014. Discovering structure in high-dimensional data through correlation explanation. In *Proc. of NIPS*.

Vinyals, O., and Le, Q. 2015. A neural conversational model. *ICML, Deep Learning Workshop*.

Walters-Williams, J., and Li, Y. 2009. Estimation of mutual information: A survey. In *International Conference on Rough Sets and Knowledge Technology*. Springer.

Wang, J.; Zhang, T.; Sebe, N.; Shen, H.; et al. 2017. A survey on learning to hash. *IEEE TPAMI*.

Wasserman, L. 2006. *All of nonparametric statistics*. Springer Science & Business Media.

Watanabe, S. 1960. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*.

Wen, T.; Vandyke, D.; Mrksic, N.; Gasic, M.; Rojas-Barahona, L.; Su, P.; Ultes, S.; and Young, S. 2017. A network-based end-to-end trainable task-oriented dialogue system. *In Proc. of EACL*.

Wu, Y.; Wu, W.; Li, Z.; and Zhou, M. 2018. Learning matching models with weak supervision for response selection in retrieval-based chatbots. In *Proc. of ACL*.

Xing, C.; Wu, Y.; Wu, W.; Huang, Y.; and Zhou, M. 2018. Hierarchical recurrent attention network for response generation. In *Proc. of AAAI*.

Zhai, K., and Williams, J. 2014. Discovering latent structure in task-oriented dialogues. In *Prof. of ACL*.

Zhang, Y.; Galley, M.; Gao, J.; Gan, Z.; Li, X.; Brockett, C.; and Dolan, B. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Proc. of NIPS*.

Zhao, T.; Lee, K.; and Eskenazi, M. 2018. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proc. of ACL*.

Zhao, K.; Lu, H.; and Mei, J. 2014. Locality Preserving Hashing. In *Proc. of AAAI*.