MRF-Chat: Improving Dialogue with Markov Random Fields

Ishaan Grover[*], **Matthew Huggins**[*], **Cynthia Breazeal** and **Hae Won Park**
Massachusetts Institute of Technology
{igrover, hugginsm, breazeal, haewon}@mit.edu

Abstract

Recent state-of-the-art approaches in open-domain dialogue include training end-to-end deep-learning models to learn various conversational features like emotional content of response, symbolic transitions of dialogue contexts in a knowledge graph and persona of the agent and the user, among others. While neural models have shown reasonable results, modelling the cognitive processes that humans use when conversing with each other may improve the agent's quality of responses. A key element of natural conversation is to tailor one's response such that it accounts for concepts that the speaker and listener may or may not know and the contextual relevance of all prior concepts used in conversation. We show that a rich representation and explicit modeling of these psychological processes can improve predictions made by existing neural network models. In this work, we propose a novel probabilistic approach using Markov Random Fields (MRF) to augment existing deep-learning methods for improved next utterance prediction. Using human and automatic evaluations, we show that our augmentation approach significantly improves the performance of existing state-of-the-art retrieval models for open-domain conversational agents.

1 Introduction

With advances in deep learning, the natural language understanding community has seen a recent proliferation of open-domain dialogue systems such as See et al. (2019); Kulikov et al. (2019); Roller et al. (2021) as well as competitions such as the Amazon Alexa Prize (Khatri et al., 2018) and ConvAI (Burtsev et al., 2018; Dinan et al., 2019). Existing approaches can be classified into two main categories: generative models and retrieval models. While the former produce responses from a generative language model (Serban et al., 2016), retrieval models aim to select the best response from a set of candidate responses given a conversation history. This paper focuses on retrieval-based models.

Most of the prior work in retrieval-based models has focused on training end-to-end models using different architectures (eg. Key-Value Memory Networks (Miller et al., 2016), gated self-attention (Zhang et al., 2018b), poly-encoder (Humeau et al., 2019) on specific datasets to statistically learn various conversational features like emotional content of response (Rashkin et al., 2019), symbolic transitions of dialogue contexts in a knowledge graph (Moon et al., 2019) and even the persona of the agent and the user (Zhang et al., 2018a)). In natural conversations, while humans often view each other as cognitive agents, we observe that prior work has not focused on the cognitive processes that humans use when conversing with each other. We posit that explicitly modeling these cognitive processes and using these models alongside existing statistical approaches can improve state-of-the-art.

In conversational inference, the cognitive theory of mutual knowledge proposes that speakers and listeners maintain mental models of the knowledge and beliefs they share with each other to find common ground for communication (Gibbs Jr, 1987; Thomas, 1986). It follows that in a two-person conversation, each speaker maintains both *(i)* a model of their partner's knowledge and *(ii)* a model of the knowledge they have communicated to their partner. These models provide information about their mutual knowledge and help in deciding the next utterance. Further, as the conversation continues, each speaker updates their mental model as they gain new information from and provide new information to their partner. Consider the following example of a conversation between two speakers:

SP 1: Did you see the Avenger's movie? (U_1)
SP 2: Yes, I loved Thor's character in it. (U_2)
SP 1: Do you like superhero movies? (U_3)

[*]Equal Contribution

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4925–4936
November 7–11, 2021.

In U_1, Speaker 1 offers “Avengers" and “Movie" as context for the conversation. Speaker 2 observes that Speaker 1 knows about “Avengers", so they must also know the related concept “Thor". Consequently, Speaker 2 offers “Thor" as context with U_2. Now, from U_1 and U_2, Speaker 1 infers that the concept "superhero movies" has the highest “mutual knowledge" and says U_3.

Along with mutual knowledge, humans also account for contextual relevance of concepts, as conversation flows from one topic to another. That is, even though a concept may be familiar to both people at one point during the conversation, it may not remain relevant when they discuss another topic.

While the theory of mutual knowledge forms the basis of grounding in conversation and contextual relevance plays a vital role in conversations, to the best of our knowledge, there hasn't been an attempt to explicitly model these processes. Based on these theories, we propose a novel probabilistic approach using Markov Random Fields (MRF) to model mutual knowledge and contextual relevance. We augment existing deep-learning methods with our model for improved next utterance prediction. In this paper, we refer to deep-learning models as *base models* and to our algorithm as *MRF-Chat*.

Our primary contribution is an algorithm (MRF-Chat) to augment existing statistical deep-learning methods to improve the performance of conversational agents. MRF-Chat is model agnostic, easy to implement and independent of the base model. Our augmentation approach achieves strong results on human and automatic evaluations in predictions made by state-of-the-art models on two widely used datasets (PersonaChat and BlendedSkillTalk).

2 Related Work

Key-Value Memory Network (Persona-Chat dataset). Persona-Chat (Zhang et al., 2018a) is a crowd-sourced dataset of conversations where each speaker responds based on a given persona. After the dataset was collected, the authors trained and evaluated several models on the corpus. At the start of each conversation, the chosen model was conditioned on either the user's persona, the agent's persona, both personas, or neither. The best performing model was a *Key-Value Memory Network* (KV-Mem) (Miller et al., 2016) that uses attention over the dialogue history and personas to choose the best response.

Poly-encoder (ConvAI2 dataset). The ConvAI2 dataset is based on Persona-Chat and involves conversations between pairs of human speakers who are each given a persona, with the goal of getting to know one another. Recently, Humeau et al. (2019) introduced Poly-encoder architectures which use self-attention to learn context features at a global rather than token level. They showed that Poly-encoders are more accurate than Bi-encoders and faster (at test time) than Cross-encoders.

Knowledge prediction from Partial Information. For an algorithm to compute mutual knowledge, a person's knowledge of related concepts from partial information must be inferred. If a person talks about "Avengers", the algorithm should infer the probability of them also knowing "superhero". We build upon our prior work (Grover et al., 2019) where we experimentally validated a model for predicting children's vocabulary from partial information of their existing knowledge. The model made assumptions based on the psycholinguistic theory of semantic learning which states that humans learn new words by forming semantic associations with words they already know. More specifically, the model was based on the following assumption: if it is observed that a child knows a given word, the child must have learned it by forming semantic associations with words they already knew. Thus, it is likely that if a child knows a given word, they also know words semantically related to it. The steps for model construction are as follows:

- **Build Semantic Network:** Nodes of the network represent words. Edges represent relationships between words. Make pairwise comparisons between n nodes in $O(n^2)$ and add an edge between two nodes if the cosine similarity between their word embeddings (Pennington et al., 2014) is above a threshold ϵ.
- **Construct corresponding MRF:** Nodes of the MRF represent the probability of knowing concepts and the pairwise potential functions represent how each node influences its neighbors (further explained in Section 4).
- **Inference:** Use existing knowledge (words) as evidence and perform inference on MRF to find conditional marginal probabilities of all the nodes in the graph.

Thus, we can find the probability of a person knowing any target concept given their knowledge about some concepts.

3 Preliminaries

A Markov Random Field (MRF) is an undirected graphical model of a joint distribution, specified by a graph $G = (V, E)$ and a set of random variables $X = \{X_1, X_2, X_3...X_n\}$ corresponding to vertices $V = \{v_1, v_2, v_3...v_n\}$. An edge e_{ij} between nodes X_i and X_j captures dependencies between nodes. These dependencies are represented by potential functions $\phi(\mathbf{x})$. Potential functions may be defined over pairs or cliques of nodes. When they are defined for pairs of nodes, the MRF is called a pairwise MRF. When $\phi_c(\mathbf{x}_c) > 0$, the probability distribution can also be expressed by a corresponding Gibbs field. For a given MRF:

$$P(X_1, X_2...X_n) = \frac{1}{Z} \prod_C \phi_c(\mathbf{x}_c) \quad (1)$$

$$Z = \sum_{\mathbf{x}} \prod_C \phi_c(\mathbf{x}_c) \quad (2)$$

$$\phi_c(\mathbf{x}_c) = e^{-E(\mathbf{x}_c)} \quad (3)$$

where C is the set of all maximal cliques, $\phi_c(\mathbf{x}_c)$ is the potential function for clique c, $E(\mathbf{x}_c)$ is the energy function for clique c, and Z is the partition function. A configuration with higher energy will have lower probability and vice-versa.

Inference on MRF gives marginal probabilities of each node. While exact inference on MRFs is computationally intractable, approximate inference algorithms such as Belief Propagation and Markov Chain Monte Carlo are often used in practice. In this paper, we use sum-product belief propagation.

4 MRF-Chat

We consider the setting where a user and conversational agent take turns interacting. We wish to incorporate the following features for the agent.

- **P1:** The agent should account for mutual knowledge. The agent should select a response utterance (from a set of candidate utterances) such that both the agent and the user maximally know about the concepts used in those utterances (common ground).
- **P2:** The agent should account for contextual relevance of concepts used in the conversation at any given time. That is, the agent should appropriately discount the mutual knowledge of a concept if it is not relevant to the current conversation (even if it was relevant earlier).

More formally, for a prior agent utterance U_{agent} and a prior user utterance U_{user}, a set of candidate response utterances $U_{candidates}$ and a base model B, we are interested in selecting a response utterance $U_{response} \in U_{candidates}$ (nomenclature included in appendix A) for the agent such that it satisfies **P1** and **P2**. We now discuss the steps to incorporate P1 and P2 separately and then discuss a method to combine them to generate a response.

4.1 P1: Mutual Knowledge

We define mutual knowledge of a concept as the probability that both the agent and user know the concept, given the concepts they have used in their respective utterances.

4.1.1 Concept Extraction

Utterance Concepts. The first step in processing utterances is to extract relevant concepts. For example, given the utterance "I love pets", the concepts "love" and "pets" should be identified. For concept extraction, we use Yake (Campos et al., 2020), which is an open-source keyword extraction tool that provides state-of-the-art performance. Given an utterance, Yake returns a list of keywords, each with a corresponding relevance score $r^{(c)} \in [0, 1]$ for concept c (since Yake scores closer to 0 indicate higher relevance, we use $r^{(c)}_{yake} = 1 - r^{(c)}$ instead). Thus, using the concept extraction module, we obtain set of extracted concepts *(i)* $C^{user}_{utterance}$ from U_{user}, *(ii)* $C^{agent}_{utterance}$ from U_{agent}, *(iii)* $C^{candidates}_{utterance}$ from $U_{candidates}$. Further, $C_{utterance} = C^{agent}_{utterance} \cup C^{user}_{utterance} \cup C^{candidates}_{utterance}$.

We note that there exist many strategies to generate the set $U_{candidates}$ and any reasonable strategy is suitable. Since our task is to improve a base model with MRF-Chat, in our experiments we use top k responses from the base model to form our candidate set. This is done for computational efficiency in running our experiments. Increasing the value of k allows for more candidates to be considered, but at the cost of increased latency.

Related Concepts. A common measure of semantic distance between two words is the cosine distance between their word embeddings. We define two words with vector representations v_1 and v_2 to be semantically related if $cos(v_1, v_2) \geq \epsilon$. For each concept in $C_{utterance}$, we find semantically related concepts in the common crawl vocabulary to obtain the set of related concepts $C_{related}$. Let the set of all concepts $C = C_{utterance} \cup C_{related}$. Since these concepts are used to build

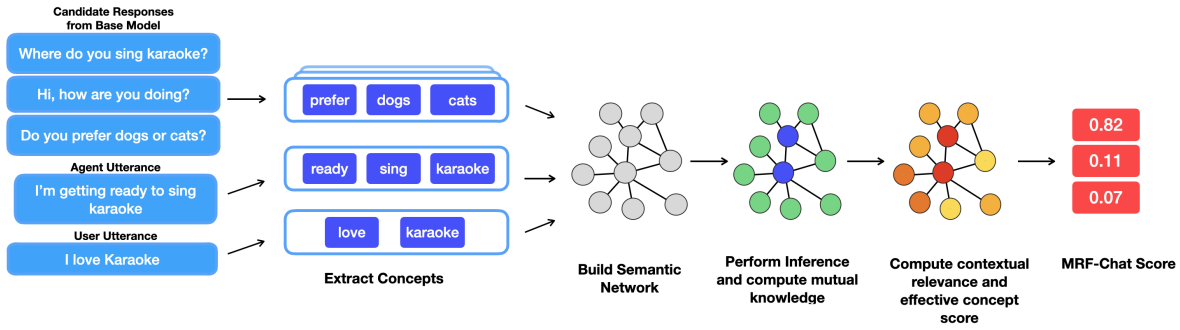

Figure 1: MRF-Chat pipeline for next utterance prediction.

a semantic graph (explained below) and represent real-world knowledge, we exclude very frequent words such as "yes", "me", "what", etc.

4.1.2 Concepts to Semantic Network

The core component of MRF-Chat is a semantic graph, where the nodes represent individual concepts and edges represent semantic relationships between them. Similar to Grover et al. (2019) (see section 2), we build a semantic network $G_{semantics} = (V_{semantics}, E_{semantics})$ from C by making pairwise comparisons[1] between word embeddings of all concepts in C.

From $G_{semantics} = (V_{semantics}, E_{semantics})$, we construct a corresponding factor graph $G_{mrf} = (X, F, E_{mrf})$ where X represents variable nodes corresponding to nodes in $G_{semantics}$, F are the factor nodes and E_{mrf} are the edges. The factors are set to the same potential functions as (Grover et al., 2019) to capture the assumptions of the psycholinguistic theory of semantic learning (see Section 2).

$$\phi(X_i, X_j) = \begin{bmatrix} e^{-(1-s(w_i,w_j))} & e^{-s(w_i,w_j)} \\ e^{-s(w_i,w_j)} & e^{-(1-s(w_i,w_j))} \end{bmatrix} \quad (4)$$

where $s(w_i, w_j)$ is the cosine distance between the word embeddings corresponding to w_i and w_j. For some concept c, $X^{(c)}$ is a Bernoulli random variable that represents the probability of the user (or agent) knowing a particular concept.

4.1.3 Inference

Let $X^{user}_{utterance}$ contain random variables corresponding to the concepts $C^{user}_{utterance}$ and $X^{agent}_{utterance}$ contain random variables corresponding to the concepts $C^{agent}_{utterance}$. Further, let $X^{(c)}_{user}$ and $X^{(c)}_{agent}$ represent the probability of the user and agent knowing a given concept c respectively. We first perform inference on G_{mrf} using variables in $X^{user}_{utterance}$ as evidence (since the user used these concepts in their utterance, we can be sure they know these concepts). This gives the conditional marginal probability of the user knowing any given concept in the graph. That is, for any concept c, we have $P(X^{(c)}_{user}|X^{user}_{utterance})$. Similarly, we perform inference on G_{mrf} using variables in $X^{agent}_{utterance}$ as evidence to obtain $P(X^{(c)}_{agent}|X^{agent}_{utterance})$. We wish to find $P(X^{(c)}_{user}, X^{(c)}_{agent}|X^{user}_{utterance}, X^{agent}_{utterance})$. Since the agent's and user's knowledge of a concept are independent (i.e, they generate utterances based on their own knowledge in a given agent-user utterance pair), $X^{(c)}_{user} \perp\!\!\!\perp X^{(c)}_{agent}$. Further, $X^{(c)}_{user}$ is independent of all variables in $X^{agent}_{utterance}$ and $X^{(c)}_{agent}$ is independent of all variables in $X^{user}_{utterance}$. Thus, we have the joint distribution,

$$P(X^{(c)}_{user}, X^{(c)}_{agent}|X^{user}_{utterance}, X^{agent}_{utterance}) = P(X^{(c)}_{user}|X^{user}_{utterance})P(X^{(c)}_{agent}|X^{agent}_{utterance}) \quad (5)$$

We now define a Bernoulli random variable $X^{(c)}_{mutual}$, representing the probability that both the user and agent know c (*Mutual knowledge*).

4.2 P2: Contextual Relevance

As an agent and user converse, each concept's relevance varies with time. For example, if the user and agent discuss "superheroes" initially but then talk about their desserts, the contextual relevance of "superheroes" decreases with time (number of turns). To capture this notion of relevance in time, we define contextual relevance of a concept as a mixture of distributions of all previous $X^{(c)}_{mutual}$ from the MRF where the weight for each distribution is exponentially decayed with every turn of the

[1] We optimize this $O(n^2)$ operation by caching previously computed distances and only computing new concepts' scores.

conversation[2]. Let mutual knowledge of concept c after the i^{th} turn-pair be $X^{(c)(i)}_{mutual}$ and let $R^{(c)}_n$ be a random variable representing the contextual relevance of c after the n^{th} turn. Then,

$$P(R^{(c)}_n) = \frac{1}{Z}\sum_{i=1}^{n} \lambda^{n-i} P(X^{(c)(i)}_{mutual}) \quad (6)$$

where Z is the normalizing constant and $\lambda \in [0, 1]$ is the rate of decay. A higher λ lowers the rate of decay and results in weighting prior mutual knowledge more heavily. While mutual knowledge is computed each turn-pair after performing inference on the MRF, prior relevance of concepts is taken into account through contextual relevance.

4.3 Next Utterance Probabilities

Given contextual relevance of each concept, we want to find the probability of each candidate utterance $u \in U_{candidates}$ being a salient next utterance.

4.3.1 Effective Concept Scores

For a given candidate response with extracted concepts, it is not only essential to reward the presence of concepts that are believed to be shared knowledge between the agent and user, but also to penalize the presence of those that are believed to not be shared. Likewise, concepts that are believed to be neutral, i.e. neither more nor less relevant than all other concepts, should have no effect.

For each concept c with contextual relevance $R^{(c)}_n$ in the n^{th} turn, we find the expected contextual relevance $E[R^{(c)}_n]$. We also find the mean expected contextual relevance, μ of all concepts in C (or nodes $V_{semantics}$). The effective concept score $S^{(c)}_n$ for concept c in the n^{th} turn is then:

$$S^{(c)}_n = E[R^{(c)}_n] - \mu \quad (7)$$

4.3.2 Utterance Score

Given a candidate utterance $u \in U_{candidates}$ in the n^{th} turn, a set of concepts $c_1, c_2, c_3 ... c_m$, effective concept scores $S^{(c_1)}_n, S^{(c_2)}_n, ..., S^{(c_m)}_n$ and yake scores $r^{(c_1)}_{yake}, r^{(c_2)}_{yake}, ..., r^{(c_m)}_{yake}$, the final score for the utterance u_{score} is given by:

$$u_{score} = \frac{1}{m}\sum_{i=1}^{m} r^{(c_i)}_{yake} S^{(c_{(i)})}_n \quad (8)$$

[2]We index from the 0^{th} turn, with the agent starting the conversation. The first inference occurs after the first turn pair.

The score of an utterance is the average effective concept score weighted by each concept's relevance in the utterance. Given scores for each utterance, we find the probability of an utterance being the next utterance according to MRF-Chat by applying softmax normalization to the scores.

4.4 Augmenting with MRF-Chat

We wish to estimate the probability of a response u being the next salient utterance and have two separate models, MRF-Chat (as described previously in this section) and Base model (deep-learning model) that estimate this probability. Thus, we have $P(u|MRF-Chat)$ and $P(u|Base)$ and want to find $P(u|MRF-Chat, Base)$. Assuming the two models to be conditionally independent, the bayes optimal method to combine the distributions is given by (Bailer-Jones and Smith, 2011) and has been used to solve other problems in machine learning (Grover et al., 2019; Griffith et al., 2013; Littman et al., 2002):

$$P(u|MRF\text{-}Chat, Base) \propto P(u|MRF\text{-}Chat)P(u|Base) \quad (9)$$

The independence assumption is reasonable because: for a given utterance u, MRF-Chat does not depend upon the Base model to compute the probability of it being the next utterance (see appendix B). The final response is the utterance with the maximum posterior.

5 Experiments

For our experiments, we compare the performance of state-of-the-art models in 2 settings: *(i)* base models augmented with MRF-Chat (Base+MRF) and *(ii)* base models alone (Base)[3].

Baseline Models. We used Poly-encoders and KV-Mem as the current state-of-the-art baselines. Humeau et al. (2019) show that poly-encoders outperform bi-encoders and all models submitted to the ConvAI2 competition including Transfer-Transfo (Wolf et al., 2019), obtaining new state-of-the-art. Additionally, poly-encoders are based on BERT pretraining. Hence, we believe that poly-encoders provide not only a reasonable but also a challenging baseline to improve upon. Further,

[3]To convert each model's output to a probability distribution over the candidates, we apply softmax before combining with MRF-Chat.

Key-Value Memory Networks are a common standard baseline used in prior work.

Datasets. *(i)* KV Memory: pre-trained on Persona-Chat (10,907 dialogues) and evaluated on Persona-Chat validation (1000 dialogues) and test (968 dialogues) sets. *(ii)* Poly-encoder: pre-trained on the ConvAI2 dataset and evaluated on validation (1,009 conversations) and test sets (980 conversations) from the BlendedSkillTalk dataset (BST) (Smith et al., 2020). We do not use the ConvAI2 test set because it is not publicly available, and the validation set alone does not have enough conversations to statistically evaluate our model. BST has twice as many conversations for evaluation and was specifically collected to evaluate models on their ability to be engaging, knowledgeable and empathetic as opposed to a single metric that previous datasets targeted. We believe this dataset provides an independent and robust test bed to evaluate our model's performance.

Multi-turn conversations. For each conversation in the dataset, we used the first 4 turns as context, processing each utterance as described in Section 4 to update marginal probabilities and contextual relevance after each turn-pair. We then produced a response to follow as the next utterance using both, the base model+MRF-Chat and the base model alone. We repeated the same experiments using the first 6 turns as context to evaluate our model with increasing conversation length.

Sensitivity to λ. We also repeat this response selection process for each value of $\lambda \in \{0, 0.3, 0.6\}$ ($\lambda = 0$ means that only the most recent user-agent utterance pair is used, ignoring previous turns). We exclude conversations in which MRF-Chat and the base model select the same response.

Hyperparameters. We used top $K = 10$ candidate responses from the base model. For building the semantic network, we used pre-trained common crawl GloVe word embeddings (300 dimensional) (Pennington et al., 2014) with a threshold $\epsilon \geq 0.6$ for adding an edge between nodes, only considering the 100,000 most common words. To exclude frequent words, we use word frequencies from the SUBTLEX-US database (Brysbaert and New, 2009), excluding words with a Zipf value of greater than 5.75 based on empirical observation.

Runtime considerations. In each turn, graph augmentation runs in $O(N^2 + NR)$ where N is the number of new and related concepts in the current turn and R is the number of prior related concepts. There is also an added cost for performing inference. However, in our experiments, we found this runtime to be computationally acceptable and suitable for deployment in real-time systems.

5.1 Evaluation Methods

Human Evaluation. We ran crowdsourcing tasks on Amazon Mechanical Turk. For each conversation, workers compared responses from the base model with MRF-Chat against the base model alone. If two values of λ produce the same response for a conversation, the response is rated only once to avoid redundancy. Inspired by Acute-Eval (Li et al., 2019), we chose questions with the highest inter-rater reliability. Workers were asked which response is better, based on the conversation, and which is more on-topic. For both, we a use four-point scale of "Response 1 is much better", "Response 1 is slightly better", "Response 2 is slightly better", and "Response 2 is much better" (see appendix C, figure 2). To determine significance, we perform a binomial test on the human ratings assuming both models perform equally well as the null hypothesis.

Automatic Evaluation. For automatic evaluation, we construct candidate sets with a ratio of 1:19 between correct and incorrect responses for each conversation (as done in the ConvAI2 competition (Dinan et al., 2019) and Humeau et al. (2019); Zhang et al. (2018a)). We report Hits@1 and Mean Reciprocal Rank(MRR) for Base+MRF-Chat and the base model alone for all values of λ and varying conversation lengths. Since we claim that MRF-Chat improves predictions of state-of-the-art, we present results on all conversations where Base+MRF-Chat gives a different next utterance response from Base alone.

5.2 Results

5.2.1 Human Evaluation

KV Memory. Tables 1 and 2 show a comparison of KV Memory+MRF-Chat with KV Memory alone across different values of λ. We find that KV Memory+MRF-Chat outperforms KV Memory on both the questions with statistical significance. That is, human annotators believe that KV Memory+MRF-Chat produced better and more on-topic responses ($p < 0.05$), for both lengths of multi-turn conversations across all values of λ.

Since λ is a hyperparameter in our model formulation, it is important to investigate our models

	Q1: Better Response			Q2: More On-Topic			
λ value	KV-Mem +MRF	KV-Mem	p-value	KV-Mem +MRF	KV-Mem	p-value	Different Responses
$\lambda = 0$	**335**	220	**$< .0001^{****}$**	**334**	217	**$< .0001^{****}$**	57.3%
$\lambda = 0.3$	**370**	279	**$.000206^{***}$**	**368**	276	**$.000168^{***}$**	67.04%
$\lambda = 0.6$	**364**	278	**$.000397^{***}$**	**362**	276	**$.000382^{***}$**	66.39%

Table 1: KV Memory+MRF-Chat outperforms KV Memory alone for all values of λ at conversation length=4.

	Q1: Better Response			Q2: More On-Topic			
λ value	KV-Mem +MRF	KV-Mem	p-value	KV-Mem +MRF	KV-Mem	p-value	Different Responses
$\lambda = 0$	**342**	283	**$.0102^{*}$**	**342**	280	**$.00723^{**}$**	64.56%
$\lambda = 0.3$	**341**	282	**$.0101^{*}$**	**347**	278	**$.00326^{**}$**	64.55%
$\lambda = 0.6$	**356**	274	**$.000625^{***}$**	**356**	273	**$.000539^{***}$**	65.35%

Table 2: KV Memory+MRF-Chat outperforms KV Memory alone for all values of λ at conversation length=6.

	Q1: Better Response			Q2: More On-Topic			
λ value	Poly +MRF	Poly	p-value	Poly +MRF	Poly	p-value	Different Responses
$\lambda = 0$	**119**	86	**$.0127^{*}$**	**119**	86	**$.0127^{*}$**	10.51%
$\lambda = 0.3$	**102**	91	.236	**99**	93	.359	9.89%
$\lambda = 0.6$	**104**	90	.175	**102**	91	.236	9.95%

Table 3: Poly+MRF-Chat outperforms Poly-encoder alone for all values of λ at conversation length=4.

	Q1: Better Response			Q2: More On-Topic			
λ value	Poly +MRF	Poly	p-value	Poly +MRF	Poly	p-value	Different Responses
$\lambda = 0$	**144**	120	.0785	**146**	116	**$.0366^{*}$**	14.03%
$\lambda = 0.3$	**141**	110	**$.0291^{*}$**	**137**	112	.0641	13.34%
$\lambda = 0.6$	115	**124**	.302	117	**120**	.448	12.7%

Table 4: Poly-encoder+MRF-Chat outperforms Poly-encoder alone for $\lambda \in (0.0, 0.3)$ at conversation length=6.

performance across different values of λ. For a conversation length of 4, we see an increase in p-value with an increase in λ. This observation is counter-intuitive since a higher value of λ means that we weigh the context of previous turns more heavily. We hypothesize that this effect is an artifact of shorter conversations. That is, it may be better to respond by only taking the last agent-user utterances into account when the conversation length is only 4 turns. Results on conversation length of 6 utterances (3 turns) support this hypothesis where we see a reversal of this trend. That is, with increase in the value of λ, we see a decrease in p-value for both the questions. Thus, for longer conversations, a higher value of λ might be preferred.

Poly-encoder human evaluation. Tables 3 and 4 show a similar comparison of Poly-encoder+MRF-Chat with Poly-encoder alone. For conversations with length of 4 utterances, the augmented Poly-encoder performs better than Poly-encoder alone across different values of λ on both questions. Further, we find that the results for $\lambda = 0$ are statistically significant. This result further supports our aforementioned hypothesis. How-

		Hits@1		MRR	
Conv. Length	λ	KV-Mem+MRF	KV-Mem	KV-Mem+MRF	KV-Mem
4	0	**0.239**	0.067	**0.402**	0.264
4	0.3	**0.246**	0.067	**0.402**	0.262
4	0.6	**0.237**	0.064	**0.397**	0.26
6	0	**0.228**	0.085	**0.391**	0.271
6	0.3	**0.226**	0.086	**0.389**	0.271
6	0.6	**0.23**	0.083	**0.390**	0.267

Table 5: KV-Memory+MRF-Chat outperforms KV-Memory alone in automatic evaluation.

		Hits@1		MRR	
Conv. Length	λ	Poly+MRF	Poly	Poly+MRF	Poly
4	0	**0.306**	0.153	**0.533**	0.437
4	0.3	**0.321**	0.115	**0.537**	0.41
4	0.6	**0.342**	0.118	**0.533**	0.418
6	0	**0.295**	0.125	**0.492**	0.391
6	0.3	**0.317**	0.122	**0.513**	0.398
6	0.6	**0.302**	0.139	**0.51**	0.41

Table 6: Poly-encoder+MRF-Chat outperforms Poly-encoder alone in automatic evaluation.

ever, since the p-values for $\lambda = 0.3$ and $\lambda = 0.6$ are higher, the subsequent trend is not clear. These results suggest that the choice of λ also depends on the type of conversations the agent is having with the user. While subjects in the Persona-Chat data collection were instructed to have conversations to get to know each other, subjects for BST were guided towards having a mix of engaging, knowledgeable and empathetic conversations.

For conversations with length of 6 utterances (3 turns), we see that Poly-encoder+MRF-Chat performs better than Poly-encoder alone for $\lambda = 0.0$ and $\lambda = 0.3$ (Table 4). Between the two values of λ, we find that the augmented model significantly outperforms the baseline model in selecting better responses when λ increases (Q1, $\lambda = 0.3$) and in producing more on-topic responses when λ is smaller (Q2, $\lambda = 0.0$). This result suggests that a better response may not always be exactly about the topic of discussion. Instead, the augmented model may introduce more semantically related and relevant concepts when $\lambda = 0.3$, making the responses better overall but less on-topic.

We also find that the augmented Poly-encoder performs slightly worse at $\lambda = 0.6$ suggesting that λ may be too high for a conversation length of 6 on this dataset. We gain valuable insight for real-world applications: as the conversation length increases, λ should gradually increase for optimal performance. Further, the rate of increase depends on the type of conversation. We leave optimal tuning of λ as a learned parameter as future work.

5.2.2 Automatic Evaluation

Tables 5 and 6 show that both KV-Memory+MRF-Chat and Poly-encoder+MRF-Chat outperform KV-Memory and Poly-encoder alone on Hits@1 and MRR metrics (for all given values of λ and for both conversation history lengths of 4 and 6). The mean improvement Δ on Hits@1 for KV-Memory is 0.159 and for poly-encoders is 0.186 (across all considered λ and conversation lengths). The mean Δ for MRR for KV-Memory is 0.13 and for poly-encoders is 0.109. Further, we see that our algorithm's performance is robust to the choice of λ and conversation length in automatic metrics. This difference in sensitivity to λ can be attributed to the fact that human judgements may have low correlation with automatic evaluation metrics (Liu et al., 2016). However, both human and automatic evaluations show significant improvements when the base model is augmented with MRF-Chat.

Utterance Length and number of concepts. Tables 7 and 8 show that KV-Mem augmented with MRF-Chat produces shorter responses with

	Utterance Length		# of Concepts	
λ	KV+MRF	KV	KV+MRF	KV
0	12.51	13.55	3.90	4.51
0.3	12.50	13.51	3.88	4.51
0.6	12.51	13.55	3.89	4.51

Table 7: Comparison of Mean utterance length and mean extracted concepts between KV-Memory+MRF-Chat and KV-Memory alone for conversation length=4

	Utterance Length		# of Concepts	
λ	Poly+MRF	Poly	Poly+MRF	Poly
0	12.68	12.54	3.93	3.34
0.3	12.84	12.56	3.97	3.37
0.6	13.03	12.64	3.97	3.37

Table 8: Comparison of Mean utterance length and mean extracted concepts between Poly-encoder+MRF-Chat and Poly-encoder alone for conversation length=4

fewer concepts while Poly-encoder augmented with MRF-Chat produces longer utterances with more concepts. We also see from human and automatic evaluations that augmenting base models with MRF-Chat improves the model's quality of responses. Since the goal of MRF-Chat is to select responses with more relevant concepts, it follows that the choice of concepts used is more important than the total number of concepts. These results align with our claim that modelling mutual knowledge and contextual relevance improves performance of state-of-the-art models by selecting utterances with the most relevant concepts.

5.2.3 Success and Failure Cases

As demonstrated through human evaluations, MRF-Chat produced more on-topic responses than the base models alone. Two examples can be found in appendix C in figure 3. In both examples, MRF-Chat chooses an on-topic response when the base model's choice is not well aligned with the conversation, serving as a method for preventing off-topic responses when better candidates are present.

While our evaluations found that MRF-Chat chose more on-topic responses than the base models alone, this is not always desirable when the conversational partner is attempting to change the topic of conversation (as seen in appendix C, figure 3). We leave the task of guiding conversational topics over time to future work.

6 Conclusion

In this work, we approach open-domain dialogue through a new lens of modelling cognitive behavior with probabilistic methods. We present a novel algorithm, MRF-Chat to improve performance of retrieval-based deep-learning models without requiring the collection of new datasets or retraining. Our method is model agnostic, easy-to-implement and independent of the base model. Using human evaluations, we present statistically significant results showing that responses produced by base models augmented with MRF-Chat were rated as better and more on-topic by human annotators when compared to those produced by base models alone. Further, using automatic metrics, we show that base+MRF outperforms base alone (KV-Mem/Poly-encoder) on Hits@1 and MRR metrics for all considered values of λ across different conversation lengths. Finally, we provide a detailed analysis of the algorithm's sensitivity to the hyperparameter λ and suggest future avenues of research.

Acknowledgements

This work was supported by the IITP grant funded by the Korea government(MSIT) (No.2020-0-00842, Development of Cloud Robot Intelligence for Continual Adaptation to User Reactions in Real Service Environments) and by the National Science Foundation under Grant No. DRL-1734443. We would like to thank Pedro Colon-Hernandez, Sooyeon Jeong, Sharifa Alghowinem, Sam Spaulding and reviewers for their valuable feedback on this work.

References

C Bailer-Jones and Kester Smith. 2011. Combining probabilities. *Data Processing and Analysis Consortium (DPAS)*.

Marc Brysbaert and Boris New. 2009. Moving beyond kucera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41:977–90.

Mikhail Burtsev, Varvara Logacheva, Valentin Malykh, Iulian Vlad Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, and Yoshua Bengio. 2018. The first conversational intelligence challenge. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 25–46. Springer.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020.

Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019. The second conversational intelligence challenge (convai2).

Raymond W Gibbs Jr. 1987. Mutual knowledge and the psychology of conversational inference. *Journal of pragmatics*, 11(5):561–588.

Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems*, 26:2625–2633.

Ishaan Grover, Hae Won Park, and Cynthia Breazeal. 2019. A semantics-based model for predicting children's vocabulary. In *IJCAI*, pages 1358–1365.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.

Chandra Khatri, Anu Venkatesh, Behnam Hedayatnia, Raefer Gabriel, Ashwin Ram, and Rohit Prasad. 2018. Alexa prize—state of the art in conversational ai. *AI Magazine*, 39(3):40–55.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *INLG'19*, pages 76–87.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Michael L Littman, Greg A Keim, and Noam Shazeer. 2002. A probabilistic approach to solving crossword puzzles. *Artificial Intelligence*, 134:23–55.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP'16*.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *EMNLP'16*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *57th Annual Mtg of ACL*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP'14*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, J. Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y.-Lan Boureau, and J. Weston. 2021. Recipes for building an open-domain chatbot. In *EACL*.

A. See, Stephen Roller, Douwe Kiela, and J. Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *NAACL*.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI'16*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, J. Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *ACL*.

Gordon P Thomas. 1986. Mutual knowledge: A theoretical basis for analyzing audience. *College English*, 48(6):580–594.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur D. Szlam, Douwe Kiela, and J. Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*.

Zhuosheng Zhang, Jiangtong Li, P. Zhu, Zhao Hai, and Gongshen Liu. 2018b. Modeling multi-turn conversation with deep utterance aggregation. In *COLING*.

A Nomenclature

B	Base Model
U_{user}	Utterance from a user
U_{agent}	Utterance from an agent
$U_{candidates}$	Set of candidate utterances
$C^{user}_{utterance}$	Set of extracted concepts from user utterance
$C^{agent}_{utterance}$	Set of extracted concepts from agent utterance
$C^{candidates}_{utterance}$	Set of extracted concepts from all candidate utterances
$C_{utterance}$	Set of all extracted concepts from user, agent and candidate utterances
$C_{related}$	Set of all related concepts from $C_{utterance}$
C	Set of all concepts (related and from utterances)
$G_{semantics}$	Semantic Network
G_{mrf}	Factor graph corresponding to Semantic Network
X	Random Vector of variable nodes in factor graph
F	Factor nodes in factor graph
$s(w_i, w_j)$	The cosine distance between word embeddings corresponding to w_i and w_j
$X^{(c)}$	Bernoulli random variable representing probability of the user/agent knowing a concept c
$X^{user}_{utterance}$	Random vector containing variables that correspond to concepts from user utterance
$X^{agent}_{utterance}$	Random vector containing variables that correspond to concepts from agent utterance
$X^{(c)}_{user}$	Probability of the user knowing a given concept c
$X^{(c)}_{agent}$	Probability of the agent knowing a given concept c
$X^{(c)}_{mutual}$	Probability that both the user and agent know c
$X^{(c)(i)}_{mutual}$	Mutual knowledge of concept c in the i^{th} turn
$R^{(c)}_n$	Random variable representing the contextual relevance of c after the n^{th}
λ	Hyperparameter of MRF-Chat. Rate of weight decay in Eq. 6.
$S^{(c)}_n$	Effective concept score of concept c in the n^{th} turn
$r^{(c)}_{yake}$	$1-$ score returned by yake for a given concept in an utterance.
u_{score}	Score for an utterance computed by MRF-chat

B Eq. 9

Here, we derive Eq. 9 following (Bailer-Jones and Smith, 2011):

$$P(u|MRF-Chat, Base) = \frac{P(MRF-Chat, Base|u)P(u)}{P(MRF-chat, Base)} \tag{10}$$

Here, we assume $MRF-chat$ and B to be conditionally independent given u. For a given utterance u, MRF-chat and base model compute their probabilities independent of each other. So,

$$P(MRF-Chat, Base|u) = P(MRF-chat|u)P(Base|u) \tag{11}$$

It follows:

$$\begin{aligned} P(u|MRF-chat, Base) &= \frac{P(MRF-Chat|u)P(Base|u)P(u)}{P(MRF-chat, Base)} \\ &= \frac{P(MRF-Chat)P(Base)}{P(MRF-chat, Base)} \times \frac{P(u|MRF-Chat)P(u|Base)}{P(u)} \\ &\propto P(u|MRF-Chat)P(u|Base) \end{aligned} \tag{12}$$

C Additional Figures

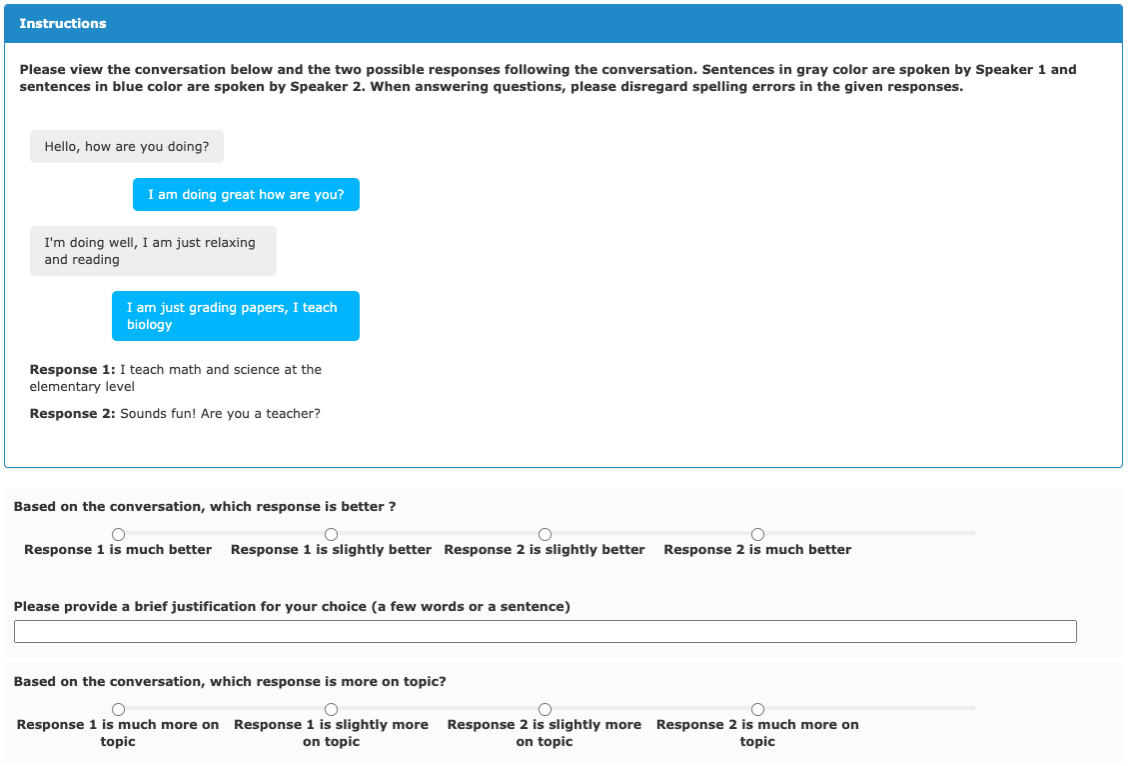

Figure 2: The Human Evaluation Setup on Mechanical Turk

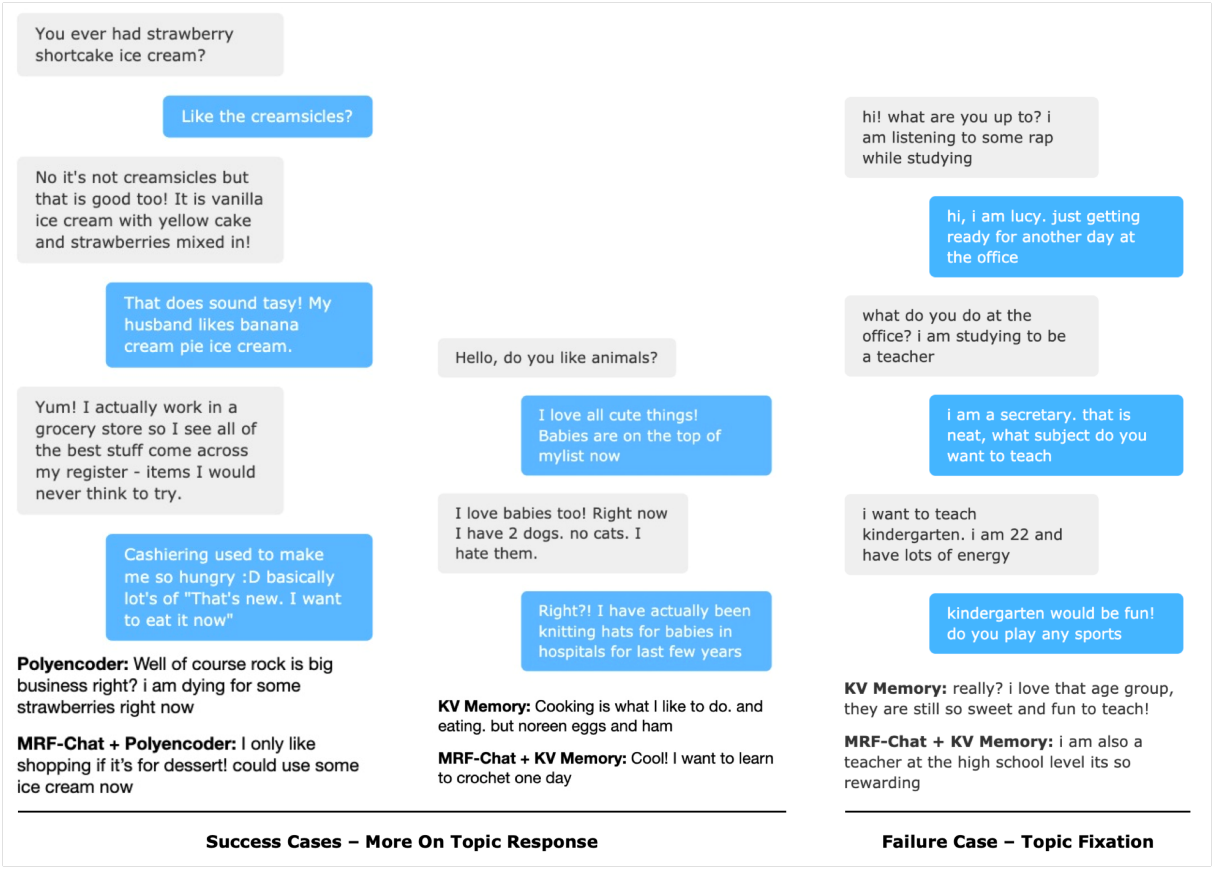

Figure 3: Example responses from Base+MRF-Chat and Base only