Knowledge Diffusion for Neural Dialogue Generation

Shuman Liu[†,§*], **Hongshen Chen**[‡], **Zhaochun Ren**[‡], **Yang Feng**[†], **Qun Liu**[◇], **Dawei Yin**[‡],
[†] Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
[‡] Data Science Lab, JD.com
[◇] ADAPT centre, School of Computing, Dublin City University
[§] University of Chinese Academy of Sciences
liushuman@ict.ac.cn, chenhongshen,renzhaochun@jd.com,
fengyang,liuqun@ict.ac.cn, yindawei@acm.org

Abstract

End-to-end neural dialogue generation has shown promising results recently, but it does not employ knowledge to guide the generation and hence tends to generate short, general, and meaningless responses. In this paper, we propose a neural knowledge diffusion (NKD) model to introduce knowledge into dialogue generation. This method can not only match the relevant facts for the input utterance but diffuse them to similar entities. With the help of facts matching and entity diffusion, the neural dialogue generation is augmented with the ability of convergent and divergent thinking over the knowledge base. Our empirical study on a real-world dataset proves that our model is capable of generating meaningful, diverse and natural responses for both factoid-questions and knowledge grounded chi-chats. The experiment results also show that our model outperforms competitive baseline models significantly.

1 Introduction

Dialogue systems are receiving more and more attention in recent years. Given previous utterances, a dialogue system aims to generate a proper response in a natural way. Compared with the traditional pipeline based dialogue system, the new method based on sequence-to-sequence model (Shang et al., 2015; Vinyals and Le, 2015; Cho et al., 2014) impressed the research communities with its elegant simplicity. Such methods are usually in an end-to-end manner: utterances are encoded by a recurrent neural network while responses are generated sequentially by another (sometimes identical) recurrent neural network. However, due to lack of universal background knowledge and common senses, the end-to-end data-driven structure inherently tends to generate meaningless and short responses, such as "haha" or "I don't know."

To bridge the gap of the common knowledge between human and computers, different kinds of knowledge bases (e.g., the freebase (Google, 2013) and DBpedia (Lehmann et al., 2017)) are leveraged. A related application of knowledge bases is question answering, where the given questions are first analyzed, followed by retrieving related facts from knowledge bases (KBs), and finally the answers are generated.The facts are usually presented in the form of "subject-relation-object" triplets, where the subject and object are entities. With the aid of knowledge triplets, neural generative question answering systems are capable of answering facts related inquiries (Yin et al., 2016; Zhu et al., 2017; He et al., 2017a), *WH* questions in particular, like "*who is Yao Ming's wife ?*".

Although answering enquiries is essential for dialogue systems, especially for task-oriented dialogue systems (Eric et al., 2017), it is still far behind a natural knowledge grounded dialogue system, which should be able to understand the facts involved in current dialogue session (so-called facts matching), as well as diffuse them to other similar entities for knowledge-based chit-chats (i.e. entity diffusion):

1) *facts matching*: in dialogue systems, matching utterances to exact facts is much harder than explicit factoid inquiries answering. Though some utterances are facts related inquiries, whose subjects and relations can be easily recognized, for some utterances, the subjects and relations are elusive, which leads the trouble in exact facts matching.

*Work done when the first author was an intern at Data Science Lab, JD.com.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 1489–1498
Melbourne, Australia, July 15 - 20, 2018.

ID	Dialogue
1	A: Who is the director of the Titanic? 泰坦尼克号的导演是谁? B: James Cameron. 詹姆斯卡梅隆。
2	A: Titanic is my favorite film! 泰坦尼克号是我最爱的电影! B: The love inside it is so touching. 里面的爱情太感人了。
3	A: Is there anything like the Titanic? 有什么像泰坦尼克号一样的电影吗? B: I think the love story in film Waterloo Bridge is beautiful, too. 我觉得魂断蓝桥中的爱情故事也很美。
4	A: Is there anything like the Titanic? 有什么像泰坦尼克号一样的电影吗? B: Poseidon is also a classic marine film. 海神号也是一部经典的海难电影。

Table 1: Examples of knowledge grounded conversations. Knowledge entities are underlined.

Table 1 shows an example: Item 1 and 2 are talking about the film "Titanic", Unlike item 1, which is a typical question answering conversation,item 2 is a knowledge related chit-chat without any explicit relation. It is difficult to define the exact fact match for item 2.

2) *entity diffusion*: another noticeable phenomenon is that the conversation usually drifts from one entity to another. In Table 1, utterances in item 3 and 4 are about entity "Titanic", however, the entity of responses are other similar films. Such entity diffusion relations are rarely captured by the current knowledge triplets. The response in item 3 shows that the two entities "Titanic" and "Waterloo Bridge" are relevant through "love stories". Item 4 suggests another similar shipwreck film of "Titanic".

To deal with the aforementioned challenges, in this paper, we propose a **n**eural **k**nowledge **d**iffusion (NKD) dialogue system to benefit the neural dialogue generation with the ability of both convergent and divergent thinking over the knowledge base, and handle factoid QA and knowledge grounded chit-chats simultaneously. NKD learns to match utterances to relevant facts; the matched facts are then diffused to similar entities; and finally, the model generates the responses with respect to all the retrieved knowledge items.

In general, our contributions are as follows:

- We identify the problem of incorporating knowledge bases and dialogue systems as facts matching and entity diffusion.
- We manage both facts matching and entity diffusion by introducing a novel knowledge diffusion mechanism and generate the responses with the retrieved knowledge items, which enable the convergent and divergent thinking over the knowledge base.
- The experimental results show that the proposed model effectively generate more diverse and meaningful responses involving more accurate relevant entities compared with the state-of-the-art baselines.

The corpus will be released upon publication.

2 Model

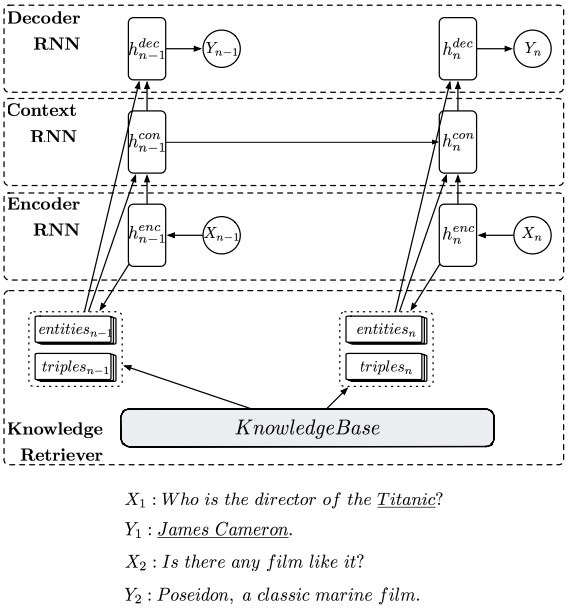

Figure 1: Neural Knowledge Diffusion Dialogue System.

Given the input utterance $X = (x_1, x_2, ..., x_{N_X})$, NKD produces a response $Y = (y_1, y_2, ..., y_{N_Y})$ containing the entities from the knowledge base K. N_X and N_Y are the number of tokens in the utterance and response respectively. The knowledge base K is a collection of knowledge facts in the form of triplets *(subject, relation, object)*. In particular, both subjects and objects are entities in this work. As illustrated in Figure 1, the model mainly consists of four components:

1. An encoder encodes the input utterance X into a vector representation.

2. A context RNN keeps the dialogue state along a conversation session. It takes the utterance representation as input, and outputs a vector guiding the response generation each turn.

3. A decoder generates the final response Y.

4. A knowledge retriever performs the facts matching and diffuses to similar entities at each turn.

Our work is built on hierarchical recurrent encoder-decoder architecture (Sordoni et al., 2015a), and a knowledge retriever network integrates the structured knowledge base into the dialogue system.

2.1 Encoder

The encoder transforms discrete tokens into vector representations. To capture information at different aspects, we learn utterance representations with two independent RNNs resulting with two hidden state sequences $H^C = (h_1^C, h_2^C, ..., h_{N_X}^C)$ and $H^K = (h_1^K, h_2^K, ..., h_{N_X}^K)$ respectively. One final hidden state $h_{N_X}^C$ is used as the input of context RNN to track the dialogue state. The other final hidden state $h_{N_X}^K$ is utilized in knowledge retriever and is designed to encode the knowledge entities and relations within the input utterances. For instance, in Figure 1, *"director"* and *"Titanic"* in X_1 are knowledge elements.

2.2 Knowledge Retriever

Knowledge retriever extracts a certain number of facts from knowledge base and specifies their importance. It enables the knowledge grounded neural dialogue system with convergent and divergent thinking ability through facts matching and entity diffusion. Figure 2 illustrates the process.

2.2.1 Facts Matching

Given the input utterance X and $h_{N_X}^K$, relevant facts are extracted from both the knowledge base and the dialogue history. A predefined number of relevant facts $F = \{f_1, f_2, ..., f_{N_f}\}$ are obtained through string matching, entity linking or named entity recognition. As shown in Figure 2, in the first sentence, *"Titanic"* is recognized as an entity, all the relevant knowledge triplets are extracted. Then, these entities and knowledge triplets are transformed into fact representations $h_f = \{h_{f_1}, h_{f_2}, ...h_{f_{N_f}}\}$ by averaging the entity embedding and relation embedding. The relevance coefficient r^f between a fact and the input utterances, ranging from 0 to 1, is calculated by a nonlinear function or a sub neural network. Here, we apply a multi-layer perceptron (MLP):

$$r_k^f = MLP([h_{N_X}^K, h_{f_k}]).$$

For the multi-turn conversation, entities in previous utterances are also inherited and reserved as depicted in Figure 2 the dotted lines. For instance, in the second sentence of Figure 2 (right one), no new fact is extracted from the input utterance. Thus it is necessary to record the history entities *"Titanic"* and *"James Cameron"*. We summarize the facts as *relevant fact representation* C^f through a weighted average of fact representations h_f:

$$C^f = \frac{\sum_{k=1}^{N_f} r_k^f h_{f_k}}{\sum_{k=1}^{N_f} r_k^f}.$$

2.2.2 Entity Diffusion

To retrieve other relevant entities, which are typically not mentioned in the dialogue utterance, we diffuse the matched facts. We calculate the similarity between the entities (except the entities that have occurred in previous utterances) in the knowledge base and the relevant fact representation through a multi-layer perceptron, resulting with a similarity coefficient r^e, ranging from 0 to 1:

$$r_k^e = MLP([h_{N_X}^K, C^f, e_k]),$$

where e_k is the entity embedding. The top N_e number of entities $E = \{e_1, e_2, ..., e_{N_e}\}$ are selected as similar entities. Then, the *similar entity representation* C^s is formalized as:

$$C^s = \frac{\sum_{k=1}^{N_e} r_k^e e_k}{\sum_{k=1}^{N_e} r_k^e}.$$

Back to the example in Figure 2, in the first turn, the matched fact of the input utterance $(Titanic, direct\_by, JamesCameron)$ is of a high relevance coefficient in "facts matching" as expected. When a fact getting matched, intuitively it is not necessary for entity diffusion. In such case, from the Figure 2, we observe that the entities in "entity diffusing" are of low similarities. In the second turn, there is no triplets matched to the utterance, while the entity *"Titanic"* achieves a much higher relevance score. Then in "entity

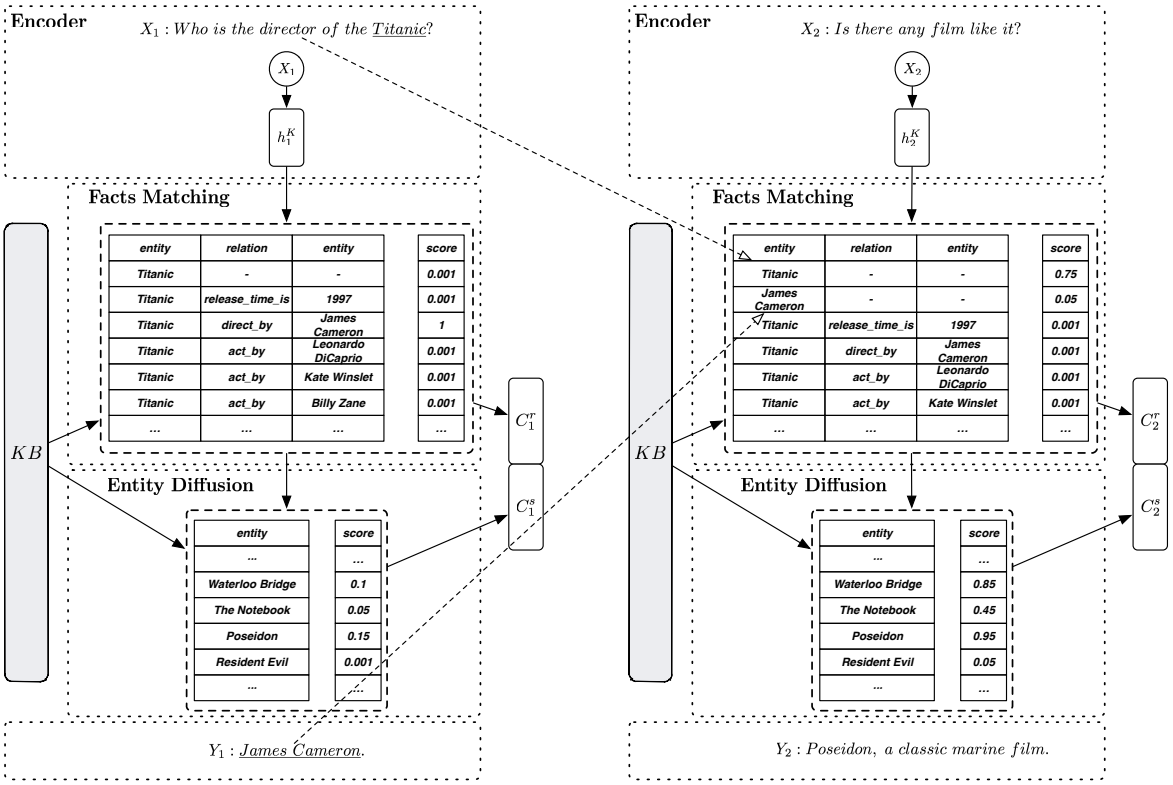

Figure 2: Knowledge Retriever. Facts related to input utterance are extracted by facts matching. Similar entities are then figured out by entity diffusion. The dotted lines show the inheritance of previous facts.

diffusion", the similar entities *"Waterloo Bridge"* and *"Poseidon"* get relatively higher similarity weights than in the first turn.

2.3 Context RNN

Context RNN records the utterance level dialogue state. It takes in the utterance representation and the knowledge representations. The hidden state of the context RNN is updated as:

$$h_t^T = RNN(h_t^C, [C^f, C^s], h_{t-1}^T).$$

h_t^T is then conveyed to the decoder to guide the response generation.

2.4 Decoder

The decoder generates the response sequentially through a word generator conditioned on h_t^T, C^f and C^s. Let C denotes the concatenation of h_t^T, C^f and C^s. Knowledge items coefficient R is the concatenation of relevance coefficient r^f and similarity coefficient r^e. We introduce two variants of word generator:

Vanilla decoder simply generates the response $Y = (y_1, y_2, ..., y_{N_y})$ according to C, R. The

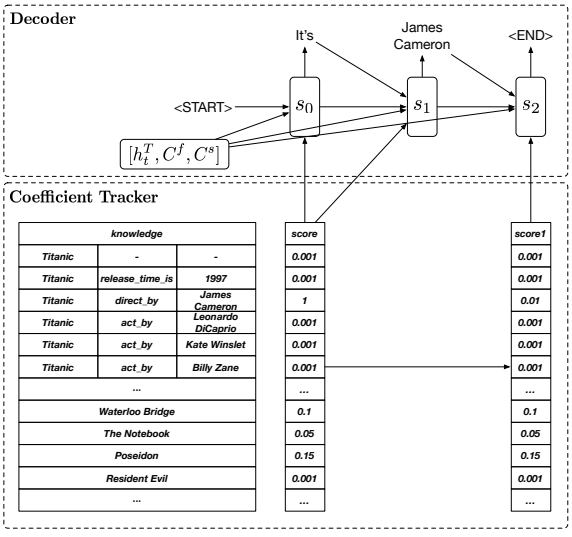

Figure 3: The decoder generates words from both vocabulary and knowledge base. A score updater keeps tracking of the knowledge item coefficients to ensure its coverage during response generation.

probability of Y is defined as

$$p(y_1, .., y_{N_y}|C, R; \theta) = p(y_1|C, R; \theta) \prod_{t=2}^{N_y} p(y_t|y_1, .., y_{t-1}, C, R; \theta),$$

where θ denotes the model parameters. The conditional probability of y_t is specified by

$$p(y_t|y_1, ..., y_{t-1}, C, R; \theta) = p(y_t|y_{t-1}, s_t, C, R; \theta),$$

where y_t is the embedding of the vocabulary or object entities of retrieved knowledge items, s_t is the decoder RNN hidden state .

Probabilistic gated decoder utilizes a gating variable z_t (Yin et al., 2016) to indicate whether the t^{th} word is generated from common vocabulary or knowledge entities. The probability of generating the t^{th} word is given by:

$$\begin{aligned} &p(y_t|y_{t-1}, s_t, C, R; \theta) \\ &=p(z_t = 0|s_t; \theta)p(y_t|y_{t-1}, s_t, C, R, z_t = 0; \theta) \\ &+p(z_t = 1|s_t; \theta)p(y_t|R, z_t = 1; \theta), \end{aligned}$$

where $p(z_t|s_t; \theta)$ is computed by a logistic regression, $p(y_t|R, z_t = 1; \theta)$ is approximated with the knowledge items coefficient R, and θ is the model parameter.

During response generation, if an entity is overused, the response diversity will be reduced. Therefore, once a knowledge item occurred in the response, the corresponding coefficient should be reduced in case that an item occurs multiple times. To keep tracking the coverage of knowledge items, we update the knowledge items coefficient R at each time step. We also explore two coverage tracking mechanisms: 1) *Mask coefficient tracker* directly reduces the coefficient of the chosen knowledge item to 0 to ensure it can never be selected as the response word again. 2) *Coefficient attenuation tracker* calculates an attenuation score i_t based on s_t, R_0, R_{t-1} and y_{t-1}:

$$i_t = DNN(s_t, y_{t-1}, R_0, R_{t-1}),$$

and then update the coefficient as:

$$R_t = i_t \cdot R_{t-1},$$

where i_t ranges from 0 to 1 to gradually decrease the coefficient.

2.5 Training

The model parameters include the embedding of vocabulary, entities, relations, and all the model components. The model is differential and can be optimized in an end-to-end manner using back-propagation. Given the training data

$$D = \{(X_1^{N_d}, Y_1^{N_d}, F_1^{N_d}, E_1^{N_d})\}$$

where N_d is the max turns of a dialogue, F denotes the set of relevant knowledge and E denotes the set of similar knowledge in response, the objective function is to minimize the negative log-likelihood:

$$\ell(D, \theta) = -\sum \sum_{i=1}^{N_D} \log p(Y_i|X_i, F_i, E_i)$$

3 Experiment

3.1 Dataset

Most existing knowledge related datasets are mainly focused on single-turn factoid question answering (Yin et al., 2016; He et al., 2017b). We here collect a multi-turn conversation corpus grounded on the knowledge base, which includes not only facts related inquiries but also knowledge-based chit-chats. The data is publicly available online[1].

We first obtain the element information of each movie, including the movie's title, publication time, directors, actors and other attributes from `https://movie.douban.com/`, a popular Chinese social network for movies. Then, entities and relations are extracted as triplets to build the knowledge base K.

To collect the question-answering dialogues, we crawled the corpus from a question-answering forum `https://zhidao.baidu.com/`. To gather the knowledge related chit-chat corpus, we mined the dataset from the social forum `https://www.douban.com/group/`. Users post their comments, feedbacks, and impressions of films and televisions on it.

The conversations are grounded on the knowledge using NER, string match, and artificial scoring and filtering rules. The statistical information of the dataset is shown in Table 2. We observed that the conversations follow the long tail distribution, where famous films and televisions are discussed repeatedly and the low rating ones are rarely mentioned.

3.2 Experiment Detail

The total 32977 conversations consisting of 104567 utterances are divided into training (32177) and testing set (800). Bi-directional LSTM (Schuster and Paliwal, 1997) is used for encoder, and the dimension of the LSTM hidden

[1]https://github.com/liushuman/neural-knowledge-diffusion

Knowledge base			Community QA	Multi-round dialogue	
#entities	#relations	#triplets	#QA pairs	#dialogues	#sentences
152568	4	766854	8121	24856	88325

Table 2: Statistics of knowledge base and conversations.

layer is set to 512. For the context RNN, the dimension of the LSTM unit is set to 1024. The dimension of word embedding shared by the vocabulary, entities and relations is also set to 512 empirically. We use Adam learning (Kingma and Ba, 2014) to update the gradient and clip the gradient in 5.0. It takes 140 to 150 epochs to train the model with a batch size of 80.

3.3 Baselines

We compare our neural knowledge diffusion model with three state-of-the-art baselines:

- **Seq2Seq**: a sequence to sequence model with vanilla RNN encoder-decoder (Shang et al., 2015; Vinyals and Le, 2015).
- **HRED**: a hierarchical recurrent encoder-decoder model.
- **GenDS**: a neural generative dialogue system that is capable of generating responses based on input message and related knowledge base (KB) (Zhu et al., 2017) .

Three variants of the neural diffusion dialogue generation model are implemented to verify different configurations of decoders.

- **NKD-ori** is the original model with a vanilla decoder and a mask coefficient tracker.
- **NKD-gated** is augmented with a probabilistic gated decoder and a mask coefficient tracker.
- **NKD-atte** utilizes a vanilla decoder and the coefficient attenuation tracker.

3.4 Evaluation Metric

Both automatic and human evaluation metrics are used to analyze the model's performance. To validate the effectiveness of facts matching and diffusion, we calculate **entity accuracy and recall** on factoid QA data set as well as the whole data set. Human evaluation rates the model in three aspects: **fluency**, **knowledge relevance** and **correctness** of the response. All these metrics range from 0 to 3, where 0 represents complete error, 1 for partially correct, 2 for almost correct, 3 for absolutely correct.

model	accuracy(%)	recall(%)
LSTM	7.8	7.5
HRED	3.7	3.9
GenDS	70.3	63.1
NKD-ori	67.0	56.2
NKD-gated	**77.6**	**77.3**
NKD-atte	55.1	46.6

Table 3: Evaluation results on factoid question answering dialogues.

model	accuracy(%)	recall(%)	entity number
LSTM	2.6	2.5	1.65
HRED	1.4	1.5	1.79
GenDS	20.9	17.4	1.34
NKD-ori	22.9	19.7	2.55
NKD-gated	**24.8**	**25.6**	1.59
NKD-atte	18.4	16.0	**3.41**

Table 4: Evaluation results on entire dataset.

3.5 Experiment Result

Table 3 displays the accuracy and recall of entities on factoid question answering dialogues. The performance of NKD is slightly better than the specific QA solution GenDS, while LSTM and HRED which are designed for chi-chat almost fail in this task. All the variants of NKD models are capable of generating entities with an accuracy of 60% to 70%, and NKD-gated achieves the best performance with an accuracy of 77.6% and a recall of 77.3%.

Table 4 lists the accuracy and recall of entities on the entire dataset including both the factoid QA and knowledge grounded chit-chats. Not surprisingly, both NKD-ori and NKD-gated outperform GenDS on the entire dataset, and the relative improvement over GenDS is even higher than the improvement in QA dialogues. It confirms that although NKD and GenDS are comparable in answering factoid questions, NKD is better at introducing the knowledge entities for knowledge grounded chit-chats.

All the NKD variants in Table 4 generate more entities than GenDS. LSTM and HRED also produce a certain amount of entities, but are of low

model	Fluency	Appropriateness of knowledge	Entire Correctness
LSTM	2.52	0.88	0.8
HRED	2.48	0.36	0.32
GenDS	**2.76**	1.36	1.34
NKD-ori	2.42	**1.92**	**1.58**
NKD-gated	2.08	1.72	1.44
NKD-atte	2.7	1.54	1.38

Table 5: Human evaluation result.

accuracies and recalls. We also noticed that NKD-gated achieves the highest accuracy and recall, but generates fewer entities compared with NKD-ori and NKD-gated, whereas NKD-atte generates more entities but also with relatively low accuracies and recalls.This demonstrates that NKD-gated not only learns to generate more entities but also maintains the quality (with a relatively high accuracy and recall).

The results of human evaluation in Table 5 also validate the superiority of the proposed model, especially on appropriateness. Responses generated by LSTM and HRED are of high fluency, but are simply repetitions, or even dull responses as "I don't know.", "Good.". NKD-gated is more adept at incorporating the knowledge base with respect to appropriateness and correctness, while NKD-atte generates more fluent responses. NKD-ori is a compromise, and obtains the best correctness in completing an entire dialogue. Four evaluators rated the scores independently. The pairwise Cohen's Kappa agreement scores are 0.67 on fluency, 0.54 on appropriateness, and 0.60 on entire correctness, which indicate a strong annotator agreement.

To our surprise, one of the variant model of NKD, which utilized both probabilistic gated decoder and coefficient attenuation tracker does not perform well on entire dataset. The accuracy of the model is quite high, but the recall is very low compared to others. We speculate that this is due to the method of minimizing negative log-likelihood during the training process, which makes the model tend to generate completely correct answers, and therefore reduces the number of generated entities.

3.6 Case Study

Table 6 shows typical examples of the generated responses. Both Item 1 and 2 are based on facts relevant utterances. NKD handles these questions by facts matching. Item 3 asks for a recommendation. NKD obtains similar entities by diffusing the entities. For item 4, 5 and 6, no explicit entity appears in the utterances. NKD is able to output appropriate recommendations through entity diffusion. The entities are recorded during the whole dialogue session, so NKD keeps recommending for several turns. Item 7 fails to generate an appropriate response because the entity in the golden response does not appear in the training set, which suggests the future work for out-of-vocabulary cases.

4 Related Work

The successes of sequence-to-sequence architecture (Cho et al., 2014; Sutskever et al., 2014) motivated investigation in dialogue systems that can effectively learn to generate a response sequence given the previous utterance sequence (Shang et al., 2015; Sordoni et al., 2015b; Vinyals and Le, 2015). The model is trained to minimize the negative log-likelihood of the training data. Despite the current progress, the lack of response diversity is a notorious problem, where the model inherently tends to generate short, general responses in spite of different inputs. Li et al. (2016a); Serban et al. (2017); Cao and Clark (2017) suggested that theses boring responses are common in training data and shorter responses are more likely to be given a higher likelihood. To tackle the problem, Li et al. (2016a) introduced a maximum mutual information training objective. Serban et al. (2017), Cao and Clark (2017) and Chen et al. (2018) used latent variables to introduce stochasticity to enhance the response diversity. Vijayakumar et al. (2016),Shao et al. (2017) and Li et al. (2016b) recognized that the greedy search decoding process, especially beam-search with a wide beam size, leads the short responses possess higher likelihoods. They reserved more diverse candidates during beam-search decoding. In this paper, we present that the absence of background knowledge and common sense is another source of lacking diversity. We augment the knowledge base to end-to-end dialogue generation.

Another research line comes from the utilizing of knowledge bases. A typical application is question-answering (QA) systems. The end-to-end QA also resort to the encoder-decoder framework (Yin et al., 2016; He et al., 2017a). Yin et al. (2016) enquired the knowledge-base to achieve one fact and answer the simple factoid questions

ID	Utterance	Response
1	X_1:Duckweed, casts 乘风破浪，演员表	Y_1:Hongda Guan Jiufeng Shi Qi Wang Zhi Cheng 关宏达 史久峰 汪潍 程之
		NKD:Hongda Guan Jiufeng Shi Qi Wang Zhi Cheng 关宏达 史久峰 汪潍 程之
2	X_1: Just spend the whole night watching the movie Silenced, a little depressed now 刚看完熔炉... 连夜看完的，略低落	Y_1:I saw it last night and had a hard time 昨晚看了，心里好难过
		NKD:The film has an atmosphere make us rethink sth 气氛有，会反思一些事
3	X_1:Which movie most representative of youth? I'd like In the Heat of the Sun, Flipped, First Love(Sing Street) 你觉得最能代表青春的电影有哪些呢 我觉得有阳光灿烂的日子，怦然心动，初恋这件小事（唱街）	Y_1:Kids Return 坏孩子的天空
		NKD:《Love Letter》《Once》 《情书》《曾经》
4	X_1:Can anyone recommend me some touching movie, tears streaming after watching 谁能给我推荐两部感人的电影，直接点的，看了可以流两缸眼泪那种	Y_1:Gift From Room 7 Hearty Paws Memories of Matsuko 7号房的礼物 人狗奇缘 被嫌弃的松子的一生
		NKD:Wedding Dress, these ones,,, 婚纱，这些，，，
5	X_1:burn brain or suspense movies please, recommend me some burn brain or suspense or horror films please! suspense films which are less popular are ok!!! 求推荐烧脑的推理悬疑电影 求各位大神推荐烧脑的悬疑恐怖电影 冷门的推理剧也爱！！！	Y_1:《The Skeleton Key》《Identity》 《万能钥匙》《致命id》
		NKD:Identity 致命id
	X_2:I've seen it, it's nice 看过哦不错呢	Y_1:《Orphan》？ 《孤儿怨》？
		NKD: Also Inception Whisper of the Heart Whisper of the Heart 还有 盗梦空间 侧耳倾听 侧耳倾听
6	X_1:It's a bit boring recently. Any good horror movies, guys? I've watched a lot. 最近有点无聊，各位豆友有什么好看的恐怖片推荐啊？看了好多了。	Y_1:Rec, A Wicked Ghost, Dead Silence, 1408 phontom horror, 死亡录像，山村老尸，死寂，1408，
		NKD:Recommend a few I've seen, 《A Wicked Ghost》, 《hide and seek》 推荐几部我看过《山村老尸》、《捉迷藏》
	X_2:I've seen it all, anything else? 卤煮都看过了。还有推荐？	Y_1:The Others, hardly call to mind 小岛惊魂，冷不丁还真想不起来
		NKD: Identity, Fight Club, Interstellar, Snowpiercer 致命id、搏击俱乐部、星际穿越、雪国列车
	X_3:Have seen it before, really great 看过，真的很不错的说	Y_3:《The Revenant》, nice movie 《亡灵》，非常好看
		NKD:So what type do you like? 那你喜欢什么类型的？
7	X_1:any awesome animation like Song of the Sea, please 求一些画风超赞的动画 类似于海洋之歌这种的，谢谢	Y_1::The Girl Without Hands 无手的少女
		NKD:strongly recommended 强烈推荐

Table 6: Examples of the generated response. Entities are underlined and Y_i denotes the gold response.

by referring to the fact. He et al. (2017a) extended this approach by augmenting the copying mechanism and enabled the output words to copy from the original input sequence. Eric et al. (2017) noticed that neural task-oriented dialogue systems often struggle to smoothly interface with a knowledge base and they addressed the problem by augmenting the end-to-end structure with a key-value retrieval mechanism where a separate attention is performed over the key of each entry in the KB. Ghazvininejad et al. (2017) represented the unstructured text as bag of words representation and also performed soft attention over the facts to retrieve a facts vector. Zhu et al. (2017) generated responses with any number of answer entities in the structured KB, even when these entities never appear in the training set. Dhingra et al. (2017) proposed a multi-turn dialogue agent which helps users search knowledge base by soft KB lookup. In our model, we perform not only facts matching to answer factoid inquiries, but also entity diffusion to infer similar entities. Given previous utterances, we retrieve the relevant facts, diffuse them, and generate responses based on diversified rele-

vant knowledge items.

5 Conclusion

In this paper, we identify the knowledge diffusion in conversations and propose an end-to-end neural knowledge diffusion model to deal with the problem. The model integrates the dialogue system with the knowledge base through both facts matching and entity diffusion, which enable the convergent and divergent thinking over the knowledge base. Under such mechanism, the factoid question answering and knowledge grounded chit-chats can be tackled together. Empirical results show the proposed model is able to generate more meaningful and diverse responses, compared with the state-of-the-art baselines. In future work, we plan to introduce reinforcement learning and knowledge base reasoning mechanisms to improve the performance.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.61662077, No.61472428). We also would like to thank all the reviewers for their insightful and valuable comments and suggestions.

References

Kris Cao and Stephen Clark. 2017. Latent variable dialogue models and their diversity. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 182–187, Valencia, Spain. Association for Computational Linguistics.

Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin. 2018. Hierarchical variational memory network for dialogue generation. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1653–1662. International World Wide Web Conferences Steering Committee.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49. Association for Computational Linguistics.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model. *arXiv preprint arXiv:1702.01932*.

Google. 2013. Freebase data dumps.

Shizhu He, Cao Liu, Kang Liu, Jun Zhao, Shizhu He, Cao Liu, Kang Liu, and Jun Zhao. 2017a. Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning. In *Meeting of the Association for Computational Linguistics*, pages 199–208.

Wei He, Kai Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2017b. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv eprint arXiv:1711.05073*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2017. Dbpedia – a large-scale, multilingual knowledge base extracted from wikipedia.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Jurafsky Dan. 2016b. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

M. Schuster and K. K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI Conference on Artificial Intelligence*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, and Brian Strope. 2017. Generating long and diverse responses with neural conversation models. *arXiv preprint arXiv:1701.03185*.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015a. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 553–562. ACM.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015b. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. 2016. Neural generative question answering. In *International Joint Conference on Artificial Intelligence*, pages 2972–2978.

Wenya Zhu, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. 2017. Flexible end-to-end dialogue system for knowledge grounded conversation. *arXiv eprint arXiv:1709.04264*.