Two are Better than One:
An Ensemble of Retrieval- and Generation-Based Dialog Systems

Yiping Song,[1] *Rui Yan,*[2] *Xiang Li,*[1] *Dongyan Zhao,*[2] *Ming Zhang*[1]
[1] School of EECS, Peking University, China
[2] Institute of Computer Science and Technology, Peking University, China
{songyiping, ruiyan, lixiang.eecs, zhaody, mzhang_cs }@pku.edu.cn

Abstract

Open-domain human-computer conversation has attracted much attention in the field of NLP. Contrary to rule- or template-based domain-specific dialog systems, open-domain conversation usually requires data-driven approaches, which can be roughly divided into two categories: retrieval-based and generation-based systems. Retrieval systems search a user-issued utterance (called a *query*) in a large database, and return a reply that best matches the query. Generative approaches, typically based on recurrent neural networks (RNNs), can synthesize new replies, but they suffer from the problem of generating short, meaningless utterances. In this paper, we propose a novel ensemble of retrieval-based and generation-based dialog systems in the open domain. In our approach, the retrieved candidate, in addition to the original query, is fed to an RNN-based reply generator, so that the neural model is aware of more information. The generated reply is then fed back as a new candidate for post-reranking. Experimental results show that such ensemble outperforms each single part of it by a large margin.

1 Introduction

Automatic dialog/conversation systems have served humans for a long time in various fields, ranging from train routing [2] to museum guiding [4]. In the above scenarios, the dialogs are domain-specific, and a typical approach to such in-domain systems is by human engineering, for example, using manually constructed ontologies [23], natural language templates [20], and even predefined dialog states [25].

Recently, researchers have paid increasing attention to open-domain, chatbot-style human-computer conversation, because of its important commercial applications, and because it tackles the real challenges of natural language understanding and generation [6, 16, 18]. For open-domain dialogs, rules and temples would probably fail as we can hardly handle the great diversity of dialog topics and natural language sentences. With the increasing number of human-human conversation utterances available on the Internet, previous studies have developed data-oriented approaches in the open domain, which can be roughly categorized into two groups: retrieval systems and generative systems.

When a user issues an utterance (called a *query*), retrieval systems search for a most similar query in a massive database (which consists of large numbers of query-reply pairs), and respond to the user with the corresponding reply [6, 7]. Through information retrieval, however, we cannot obtain new utterances, that is, all replies have to appear in the database. Also, the ranking of candidate replies is usually judged by surface forms (e.g., word overlaps, *tf·idf* features) and hardly addresses the real semantics of natural languages.

Generative dialog systems, on the other hand, can synthesize a new sentence as the reply by language models [16, 18, 19]. Typically, a recurrent neural network (RNN) captures the query's semantics with one or a few distributed, real-valued vectors (also known as *embeddings*); another RNN decodes the query embeddings to a reply. Deep neural networks allow complicated interaction by multiple non-linear transformations; RNNs are further suitable for modeling time-series data (e.g., a sequence of words) especially when enhanced with long
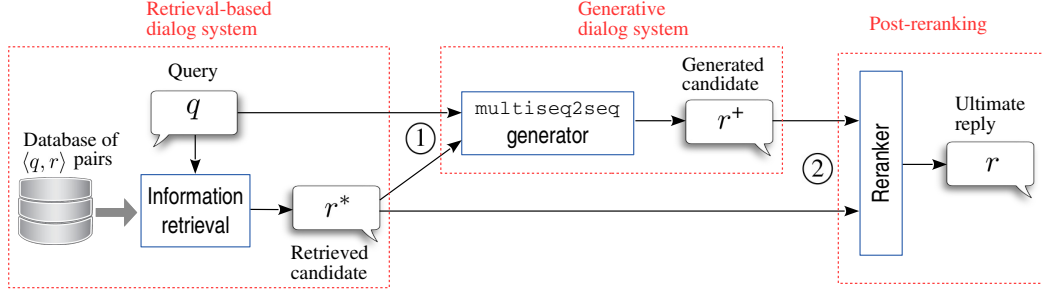
Figure 1: The overall architecture of our model ensemble. We combine retrieval and generative dialog systems by ① enhancing the generator with the retrieved candidate and by ② post-reranking of both retrieved and generated candidates.

short term memory (LSTM) or gated recurrent units (GRUs). Despite these, RNN also has its own weakness when applied to dialog systems: the generated sentence tends to be short, universal, and meaningless, for example, "I don't know" [8] or "something" [16]. This is probably because chatbot-like dialogs are highly diversified and a query may not convey sufficient information for the reply. Even though such universal utterances may be suited in certain dialog context, they make users feel boring and lose interest, and thus are not desirable in real applications.

In this paper, we are curious if we can combine the above two streams of approaches for open-domain conversation. To this end, we propose an ensemble of retrieval and generative dialog systems. Given a user-issued query, we first obtain a candidate reply by information retrieval from a large database. The query, along with the candidate reply, is then fed to an utterance generator based on the "bi-sequence to sequence" (biseq2seq) model [30]. Such sequence generator takes into consideration the information contained in not only the query but also the retrieved reply; hence, it alleviates the low-substance problem and can synthesize replies that are more meaningful. After that we use the scorer in the retrieval system again for post-reranking. This step can filter out less relevant retrieved replies or meaningless generated ones. The higher ranked candidate (either retrieved or generated) is returned to the user as the reply.

From the above process, we see that the retrieval and generative systems are integrated by two mechanisms: (1) The retrieved candidate is fed to the sequence generator to mitigate the "low-substance" problem; (2) The post-reranker can make better use of both the retrieved candidate and the generated utterance. In this sense, we call our overall approach an *ensemble* in this paper. To the best of our knowledge, we are the first to combine retrieval and generative models for open-domain conversation.

Experimental results show that our ensemble model consistently outperforms each single component in terms of several subjective and objective metrics, and that both retrieval and generative methods contribute an important portion to the overall approach. This also verifies the rationale for building model ensembles for dialog systems.

2 The Proposed Model Ensemble

2.1 Overview

Figure 1 depicts the overall framework of our proposed ensemble of retrieval and generative dialog systems. It mainly consists of the following components.

- When a user sends a query utterance q, our approach utilizes a state-of-the-practice information retrieval system to search for a query-reply pair $\langle q^*, r^* \rangle$ that best matches the user-issued query q. The corresponding r^* is retrieved as a candidate reply.

- Then a biseq2seq model takes the original query q and the retrieved candidate reply r^* as input,

each sequence being transformed to a fixed-size vector. These two vectors are concatenated and linearly transformed as the initial state of the decoder, which generates a new utterance r^+ as another candidate reply.

- Finally, we use a reranker (which is a part of the retrieval system) to select either r^* or r^+ as the ultimate response to the original query q.

In the rest of this section, we describe each component in detail.

2.2 Retrieval-Based Dialog System

Information retrieval is among prevailing techniques for open-domain, chatbot-style human-computer conversation [6, 7].

We utilize a state-of-the-practice retrieval system with extensive manual engineering and on a basis of tens of millions of existing human-human utterance pairs. Basically, it works in a two-step retrieval-and-ranking strategy, similar to the `Lucene`[1] and `Solr`[2] systems.

First, a user-issued utterance is treated as bag-of-words features with stop-words being removed. After querying it in a knowledge base, we obtain a list containing up to 1000 query-reply pairs $\langle q^*, r^* \rangle$, whose queries share most words as the input query q. This step retrieves coarse-grained candidates efficiently, which is accomplished by an inversed index.

Then, we measure the relatedness between the query q and each $\langle q^*, r^* \rangle$ pair in a fine-grained fashion. In our system, both q-q^* and q-r^* relevance scores are considered. A classifier judges whether q matches q^* and r^*; its confidence degree is used as the scorer. We have tens of features, and several important ones include word overlap ratio, the cosine measure of a pretrained topic model coefficients, and the cosine measures of word embedding vectors. (Details are beyond the scope of this paper; any well-designed retrieval system might fit into our framework.)

In this way, we obtain a query-reply pair $\langle q^*, r^* \rangle$ that best matches the original query q; the corresponding utterance r^* is considered as a candidate reply retrieved from the database.

2.3 The `biseq2seq` Utterance Generator

Using neural networks to build end-to-end trainable dialog systems has become a new research trend in the past year. A generative dialog system can synthesize new utterances, which is complementary to retrieval-based methods.

Typically, an encoder-decoder architecture is applied to encode a query as vectors and to decode the vectors to a reply utterance. With recurrent neural networks (RNNs) as the encoder and decoder, such architecture is also known as a `seq2seq` model, which has wide applications in neural machine translation [21], abstractive summarization [15], etc. That being said, previous studies indicate `seq2seq` has its own shortcoming for dialog systems. [12] suggests that, in open-domain conversation systems, the query does not carry sufficient information for the reply; that the `seq2seq` model thus tends to generate short and meaningless sentences with little substance.

To address this problem, we adopt a `biseq2seq` model, which is proposed in [30] for multi-source machine translation. The `biseq2seq` model takes into consideration the retrieved reply as a reference in addition to query information (Figure 2). Hence, the generated reply can be not only fluent and logical with respect to the query, but also meaningful as it is enhanced by a retrieved candidate.

Specifically, we use an RNN with gated recurrent units (GRUs) for sequence modeling. Let $\boldsymbol{x}_t$ be the word

---

[1]http://lucene.apache.org
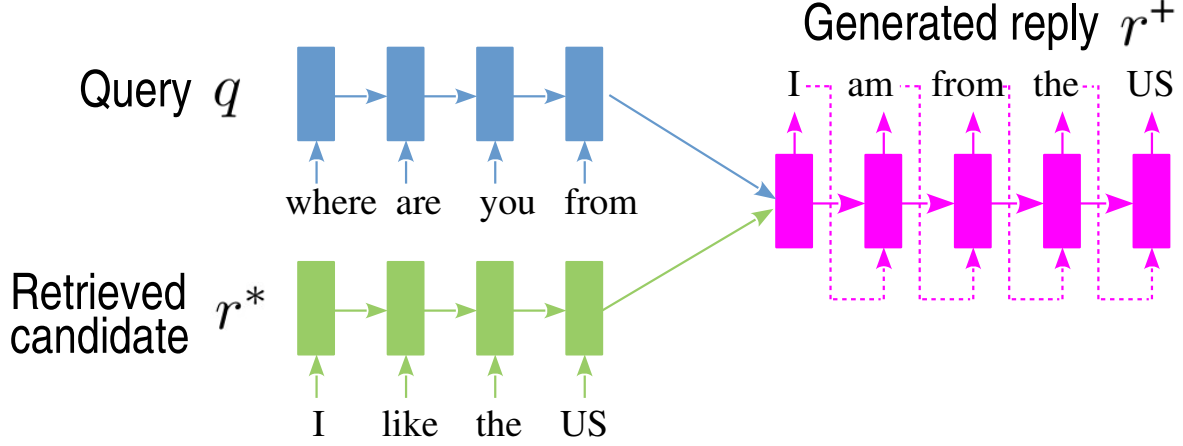[2]http://lucene.apache.org/solr

Figure 2: The `biseq2seq` model, which takes as input a query q and a retrieved candidate reply r^*; it outputs a new reply r^+.

embeddings of the time step t and $\boldsymbol{h}_{t-1}$ be the previous hidden state of RNN. We have

$$\boldsymbol{r}_t = \sigma(W_r \boldsymbol{x}_t + U_r \boldsymbol{h}_{t-1} + \boldsymbol{b}_r) \tag{1}$$

$$\boldsymbol{z}_t = \sigma(W_z \boldsymbol{x}_t + U_r \boldsymbol{h}_{t-1} + \boldsymbol{b}_z) \tag{2}$$

$$\tilde{\boldsymbol{h}}_t = \tanh\left(W_x \boldsymbol{x}_t + U_x(\boldsymbol{r}_t \circ \boldsymbol{h}_{t-1})\right) \tag{3}$$

$$\boldsymbol{h}_t = (1 - \boldsymbol{z}_t) \circ \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \circ \tilde{\boldsymbol{h}}_t \tag{4}$$

where $\boldsymbol{r}_t$ and $\boldsymbol{z}_t$ are known as gates, W's and $\boldsymbol{b}$'s are parameters, and "$\circ$" refers to element-wise product.

After two RNNs go through q and r^*, respectively, we obtain two vectors capturing their meanings. We denote them as bold letters $\boldsymbol{q}$ and $\boldsymbol{r}^*$, which are concatenated as $[\boldsymbol{q}; \boldsymbol{r}^*]$ and linearly transformed before being fed to the decoder as the initial state.

During reply generation, we also use GRU-RNN, given by Equations 1–4. But at each time step, a softmax layer outputs the probability that a word would occur in the next step, i.e.,

$$p(w_i|\boldsymbol{h}_t) = \frac{\exp\left\{W_i^\top \boldsymbol{h}_t\right\} + \boldsymbol{b}}{\sum_j \exp\left\{W_j^\top \boldsymbol{h}_t + \boldsymbol{b}\right\}} \tag{5}$$

where W_i is the i-th row of the output weight matrix (corresponding to w_i) and $\boldsymbol{b}$ is a bias term.

Notice that we assign different sets of parameters—indicated by three colors in Figure 2—for the two encoders (q and r^*) and the decoder (r^+). This treatment is because the RNNs' semantics differ significantly from one another (even between the two encoders).

2.4 Post-Reranking

Now that we have a retrieved candidate reply r^* as well as a generated one r^+, we select one as the final reply by the q-r scorer in the retrieval-based dialog system (described in previous sections and not repeated here).

Using manually engineered features, this step can eliminate either meaningless short replies that are unfortunately generated by `biseq2seq` or less relevant replies given by the retrieval system. We call this a *post-reranker* in our model ensemble.

2.5 Training

We train each component separately because the retrieval part is not end-to-end learnable.

Dataset Split	# of samples
Retrieval (Database)	7,053,820
Matching scorer (Train)	50,000
Generator (Train)	1,500,000
Validation	100,000
Testing	6,741

Table 1: Statistics of our datasets.

In the retrieval system, we use the classifier's confidence as the relevance score. The training set consists of 10k samples, which are either in the original human-human utterance pairs or generated by negative sampling. We made efforts to collect binary labels from a crowd-sourcing platform, indicating whether a query is relevant to another query and whether it is relevant to a particular reply. We find using crowd-sourced labels results in better performance than original negative sampling.

For `biseq2seq`, we use human-human utterance pairs $\langle q, r \rangle$ as data samples. A retrieved candidate r^* is also provided as the input when we train the neural network. Standard cross-entropy loss of all words in the reply is applied as the training objective. For a particular training sample whose reply is of length T, the cost is

$$J = -\sum_{i=1}^{T}\sum_{j=1}^{V} t_j^{(i)} \log y_j^{(i)} \tag{6}$$

where $\boldsymbol{t}^{(i)}$ is the one-hot vector of the next target word, serving as the groundtruth, $\boldsymbol{y}$ is the output probability by softmax, and V is the vocabulary size. We adopt mini-batched AdaDelta [29] for optimization.

3 Evaluation

In this section, we evaluate our model ensemble on Chinese (language) human-computer conversation. We first describe the datasets and settings. Then we compare our approach with strong baselines.

3.1 Experimental Setup

Typically, a very large database of query-reply pairs is a premise for a successful retrieval-based conversation system, because the reply must appear in the database. For RNN-based sequence generators, however, it is time-consuming to train with such a large dataset; RNN's performance may also saturate when we have several million samples.

To construct a database for information retrieval, we collected human-human utterances from massive online forums, microblogs, and question-answering communities, such as Sina Weibo,[3] Baidu Zhidao,[4] and Baidu Tieba.[5] We filtered out short and meaningless replies like "..." and "Errr." In total, the database contains 7 million query-reply pairs for retrieval.

For the generation part, we constructed another dataset from various resources in public websites comprising 1,606,741 query-reply pairs. For each query q, we searched for a candidate reply r^* by the retrieval component and obtained a tuple $\langle q, r^*, r \rangle$. As a friendly reminder, q and r^* are the input of `biseq2seq`, whose output should approximate r. We randomly selected 100k triples for validation and another 6,741 for testing. The train-val-test split remains the same for all competing models.

The `biseq2seq` then degrades to an utterance autoencoder [9]. Also, the validation and test sets are disjoint with the training set and the database for retrieval, which complies with the convention of machine learning.

---

[3] http://weibo.com
[4] http://zhidao.baidu.com
[5] http://tieba.baidu.com

Method	Human Score	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Retrieval	0.996	5.707	3.092	2.406	**2.094**
seq2seq	0.907	3.676	1.228	0.564	0.286
biseq2seq	0.966	7.762	3.293	2.056	1.487
Rerank(Retrieval,seq2seq)	1.030	4.500	2.041	1.364	1.060
Rerank(Retrieval,biseq2seq)	**1.131**	**7.260**	**3.503**	**2.480**	2.000

Table 2: Results of our ensemble and competing methods in terms of average human scores and BLEUs. Inter-annotator agreement for human annotation: Fleiss' $\kappa = 0.2824$ [3], std $= 0.4031$, indicating moderate agreement.

To train our neural models, we followed [18] for hyperparameter settings. All embeddings were set to 620-dimensional and hidden states 1000d. We applied AdaDelta with a mini-batch size of 80 and other default hyperparameters for optimization. Chinese word segmentation was performed on all utterances. We kept a same set of 100k words (Chinese terms) for two encoders, but 30k for the decoder due to efficiency concerns. The three neural networks do not share parameters (neither connection weights nor embeddings).

We did not tune the above hyperparameters, which were set empirically. The validation set was used for early stop based on the perplexity measure.

3.2 Competing Methods

We compare our model ensemble with each individual component and provide a thorough ablation test. Listed below are the competing methods in our experiments.

- *Retrieval*. A state-of-the-practice dialog system, which is a component of our model ensemble; it is also a strong baseline because of extensive human engineering.

- *seq2seq*. An encoder-encoder framework [21], first introduced in [18] for dialog systems.

- *biseq2seq*. Another component in our approach, adapted from [30], which is essentially a *seq2seq* model extended with a retrieved reply.

- *Rerank(Retrieval,seq2seq)*. Post-reranking between a retrieved candidate and one generated by seq2seq.

- *Rerank(Retrieval,biseq2seq)*. This is the full proposed model ensemble.

All baselines were trained and tuned in a same way as our full model, when applicable, so that the comparison is fair.

3.3 Overall Performance

We evaluated our approach in terms of both subjective and objective evaluation.

- Human evaluation, albeit time- and labor-consuming, conforms to the ultimate goal of open-domain conversation systems. We asked three educated volunteers to annotate the results using a common protocol known as *pointwise annotation* [10, 12, 18]. In other words, annotators were asked to label either "0" (bad), "1" (borderline), or "2" (good) to a query-reply pair. The subjective evaluation was performed in a strict random and blind fashion to rule out human bias.

- We adopted BLEU-1, BLEU-2, BLEU-3 and BLEU-4 as automatic evaluation. While [11] further aggressively argues that no existing automatic metric is appropriate for open-domain dialogs, they show a slight positive correlation between BLEU-2 and human evaluation in non-technical Twitter domain, which is similar to our scenario. We nonetheless include BLEU scores as expedient objective evaluation, serving as supporting evidence. BLEUs are also used in [8] for model comparison and in [12] for model selection.

6

Method	Entropy	Length
seq2seq	7.420	7.362
biseq2seq	**8.302**	**8.185**
Groundtruth	8.625	12.62

Table 3: Entropy and length of generated replies. We also include groundtruth for reference. A larger entropy value indicates that the replies are less common, and probably, more meaningful.

Notice that, automatic metrics were computed on the entire test set, whereas subjective evaluation was based on 79 randomly chosen test samples due to the limitation of human resources available.

We present our main results in Table 2. As shown, the retrieval system, which our model ensemble is based on, achieves better performance than RNN-based sequence generation. The result is not consistent with [18], where their RNNs are slightly better than retrieval-based methods. After closely examining their paper, we find that their database is multiple times smaller than ours, which may, along with different features and retrieval methods, explain the phenomenon. This also verifies that the retrieval-based dialog system in our experiment is a strong baseline to compare with.

Combining the retrieval system and the RNN generator by bi-sequence input and post-reranking, we achieve the highest performance in terms of both human evaluation and BLEU scores. Concretely, our model ensemble outperforms the state-of-the-practice retrieval system by $+13.6\%$ averaged human scores, which we believe is a large margin.

3.4 Analysis and Discussion

Having verified that our model ensemble achieves better performance than all baselines, we are further curious how each gadget contributes to our final system. Specially, we focus on the following research questions.

RQ1: What is the performance of biseq2seq (the ① step in Figure 1) in comparison with traditional seq2seq?

From the BLEU scores in Table 2, we see biseq2seq significantly outperforms conventional seq2seq, showing that, if enriched with a retrieved human utterance as a candidate, the encoder-decoder framework can generate much more human-like utterances.

We then introduce in Table 3 another measure, the entropy of a sentence, defined as

$$-\frac{1}{|R|} \sum_{w \in R} \log p(w)$$

where R refers to all replies. Entropy is used in [17] and [12] to measure the serendipity of generated utterances.[6] The results in Table 3 confirm that biseq2seq indeed integrates information from the retrieved candidate, so that it alleviates the "low-substance" problem of RNNs and can generate utterances more meaningful than traditional seq2seq. And the statistic result also displays that biseq2seq generates longer sentences than seq2seq approach.

RQ2: How do the retrieval- and generation-based systems contribute to post-reranking (the ② step in Figure 1)?

We plot in Figure 3 the percentage by which the post-raranker chooses a retrieved candidate or a generated one. In the retrieval-and-seq2seq ensemble (Figure 3a), 54.65% retrieved results and 45.35% generated ones are selected. In retrieval-and-biseq2seq ensemble, the percentage becomes 44.77% vs. 55.23%. The trend further indicates that biseq2seq is better than seq2seq (at least) from the reranker's point of view. More

---

[6]Notice that, the entropy of retrieved replies is not a fair metric to compare in Table 3, because the retrieval system has filtered out short, meaningless utterances in advance by surface statistics (e.g., length). We nevertheless report the result here out of curiosity: its entropy is 9.507, which is even higher than groundtruth.
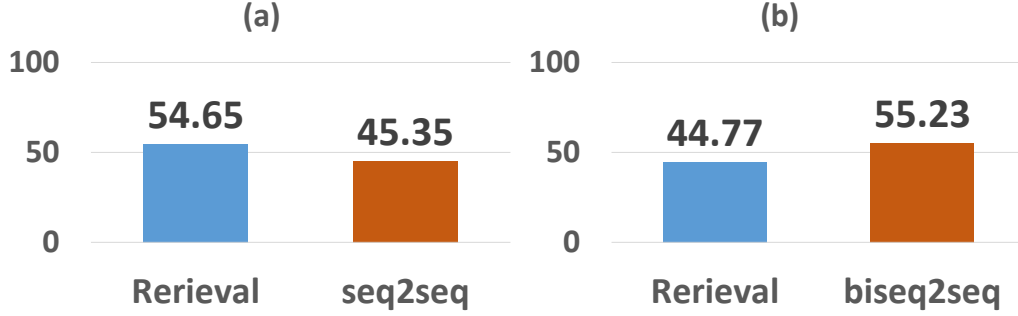
Figure 3: The percentage by which our post-reranker chooses a retrieved reply or a generated reply. (a) Ensemble of *Retrieval* and `seq2seq`; (b) Ensemble of *Retrieval* and `biseq2seq`.

	Utterance (Translated)	**Reranked**
Query	手机拍照效果不错 (This mobile phone's photo effect is pretty good.)	
Retrieved	所以好心动 (I really have a crush on it.)	
`biseq2seq`	心动不如行动 (Rushing for it rather than having a crush on it.)	←
`seq2seq`	哈哈 (Haha)	
Query	有给孕妇推荐的保健品啥的么? (Could you please recommend some health products for pregnant woman?)	
Retrieved	打算上孕妇装? (Do you want metarnity clothes?)	
`biseq2seq`	打算买护肤品么? (Do you want some skin care products?)	←
`seq2seq`	不好意思 (I'm sorry.)	

Table 4: Examples of retrieved replies and generated ones. An arrow "←" indicates the one selected during post-reranking. Also included are replies generated by conventional `seq2seq`. Notice that it is not part of our model and thus not considered for reranking.

importantly, as the percentages are close to 50%, both the retrieval system and the generation system contribute a significant portion to our final ensemble.

RQ3: Do we obtain further gain by combining the two gadgets ① and ② in Figure 1?

We would also like to verify if the combination of `biseq2seq` and post-reranking mechanisms will yield further gain in our ensemble. To test this, we compare the full model *Rerank(Retrieval,`biseq2seq`)* with an ensemble that uses traditional `seq2seq`, i.e., *Rerank(Retrieval,`seq2seq`)*. As indicated in Table 2, even with the post-reranking mechanism, the ensemble with underlying `biseq2seq` still outperforms the one with `seq2seq`. Likewise, *Rerank(Retrieval,`biseq2seq`)* outperforms both *Retrieval* and `biseq2seq`. These results are consistent in terms of all metrics except a BLEU-4 score.

Through the above ablation tests, we conclude that both gadgets (`biseq2seq` and post-reranking) play a role in our ensemble when we combine the retrieval and the generative systems.

3.5 Case Study

Table 4 presents two examples of our ensemble and its "base" models. We see that `biseq2seq` is indeed influenced by the retrieved candidates. As opposed to traditional `seq2seq`, several content words in the retrieved replies (e.g., *crush*) also appear in `biseq2seq`'s output, making the utterances more meaningful. The

post-reranker also chooses a more appropriate candidate as the reply.

4   Related Work

In early years, researchers mainly focus on domain-specific dialog systems, e.g., train routing [2], movie information [1], and human tutoring [5]. Typically, a pre-constructed ontology defines a finite set of slots and values, for example, cuisine, location, and price range in a food service dialog system; during human-computer interaction, a state tracker fills plausible values to each slot from user input, and recommend the restaurant that best meets the user's requirement [13, 24, 26].

In the open domain, however, such slot-filling approaches would probably fail because of the diversity of topics and natural language utterances. [6] applies information retrieval techniques to search for related queries and replies. [7] and [28] use both shallow hand-crafted features and deep neural networks for matching. [10] proposes a random walk-style algorithm to rank candidate replies. In addition, their model can introduce additional content (related entities in the dialog context) by searching a knowledge base when a stalemate occurs during human-computer conversation.

Generative dialog systems have recently attracted increasing attention in the NLP community. [14] formulates query-reply transformation as a phrase-based machine translation. Since the last year, the renewed prosperity of neural networks witnesses an emerging trend in using RNN for dialog systems [16, 18, 19, 22]. However, a known issue with RNN is that it prefers to generate short, meaningless utterances. [8] proposes a mutual information objective in contrast to the conventional maximum likelihood criterion. [12] and [27] introduce additional content (either the most mutually informative word or topic information) to the reply generator. [17] applies a variational encoder to capture query information as a distribution, from which a random vector is sampled for reply generation.

To the best of our knowledge, we are the first to combine retrieval-based and generation-based dialog systems. The use of `biseq2seq` and post-reranking is also a new insight of this paper.

5   Conclusion and Future Work

In this paper, we propose a novel ensemble of retrieval-based and generation-based open-domain dialog systems. The retrieval part searches a best-match candidate reply, which is, along with the original query, fed to an RNN-based `biseq2seq` reply generator. The generated utterance is fed back as a new candidate to the retrieval system for post-reranking. Experimental results show that our ensemble outperforms its underlying retrieval system and generation system by a large margin. In addition, the ablation test demonstrates both the `biseq2seq` and post-reranking mechanisms play an important role in the ensemble.

Our research also points out several promising directions for future work, for example, developing new mechanisms of combining retrieval and generative dialog systems, as well as incorporating other data-driven approaches to human-computer conversation.

References

[1] Jennifer Chu-Carroll. MIMIC: An adaptive mixed initiative spoken dialogue system for information queries. In *Proc. Conf. Applied Natural Language Processing*, pages 97–104, 2000.

[2] G. Ferguson, J. Allen, B. Miller, et al. Trains-95: Towards a mixed-initiative planning assistant. In *AIPS*, pages 70–77, 1996.

[3] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378, 1971.

[4] Nadine Glas, Ken Prepin, and Catherine Pelachaud. Engagement driven topic selection for an information-giving agent. In *Proc. Workshop on the Semantics and Pragmatics of Dialogue*, 2015.

[5] A. Graesser, P. Chipman, B. Haynes, and A. Olney. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Trans. Education*, 48(4):612–618, 2005.

[6] Charles Lee Isbell, Michael Kearns, Dave Kormann, Satinder Singh, and Peter Stone. Cobot in LambdaMOO: A social statistics agent. In *AAAI*, pages 36–41, 2000.

[7] Zongcheng Ji, Zhengdong Lu, and Hang Li. An information retrieval approach to short text conversation. *arXiv preprint arXiv:1408.6988*, 2014.

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

[9] Jiwei Li, Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. In *ACL-IJCNLP*, pages 1106–1115, 2015.

[10] Xiang Li, Lili Mou, Rui Yan, and Ming Zhang. StalemateBreaker: A proactive content-introducing approach to automatic human-computer conversation. In *IJCAI*, pages 2845–2851, 2016.

[11] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP (to appear)*, 2016.

[12] Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. *arXiv preprint arXiv:1607.00970*, 2016.

[13] Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Multi-domain dialog state tracking using recurrent neural networks. In *ACL-IJCNLP*, pages 794–799, 2015.

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

[15] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *EMNLP*, pages 379–389, 2015.

[16] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3783, 2016.

[17] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069*, 2016.

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

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

[20] Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. Continuously learning neural dialogue management. *arXiv preprint arXiv:1606.02689*, 2016.

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

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

[23] Tsung-Hsien Wen, Milica Gasic, Dongho Kim, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *SIGDIAL*, pages 275–284, 2015.

[24] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*, 2016.

[25] Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. The dialog state tracking challenge. In *SIGDIAL*, pages 404–413, 2013.

[26] Jason D Williams. Web-style ranking and SLU combination for dialog state tracking. In *SIGDIAL*, pages 282–291, 2014.

[27] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. Topic augmented neural response generation with a joint attention mechanism. *arXiv preprint arXiv:1606.08340*, 2016.

[28] Rui Yan, Yiping Song, and Hua Wu. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *SIGIR*, pages 55–64, 2016.

[29] Matthew D Zeiler. AdaDelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[30] Barret Zoph and Kevin Knight. Multi-source neural translation. In *NAACL-ACL*, pages 30–34, 2016.