Integrating Bayesian and Neural Networks for Discourse Coherence

Jinhua Peng, Zongyang Ma, Di Jiang, Hua Wu
Baidu Inc. China
{pengjinhua,mazongyang,jiangdi,wu_hua}@baidu.com

ABSTRACT

In dialogue systems, discourse coherence is an important concept that measures semantic relevance between an utterance and its context. It plays a critical role in determining the inappropriate reply of dialogue systems with regard to a given dialogue context. In this paper, we present a novel framework for evaluating discourse coherence by seamlessly integrating Bayesian and neural networks. The Bayesian network corresponds to Coherence-Pivoted Latent Dirichlet Allocation (cpLDA). cpLDA concentrates on generating the fine-grained topics from dialogue data and takes both local and global semantics into account. The neural network corresponds to Multi-Hierarchical Coherence Network (MHCN). Coupled with cpLDA, MHCN quantifies discourse coherence between an utterance and its context by comprehensively utilizing original texts, topic distribution and topic embedding. Extensive experiments show that the proposed framework yields superior performance comparing with the state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies → Discourse, dialogue and pragmatics**.

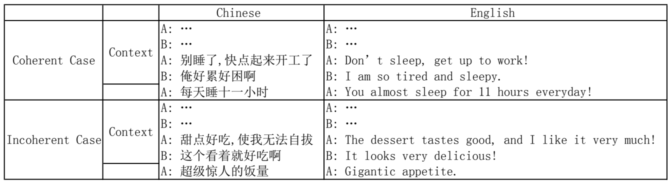
KEYWORDS

Bayesian Network, Multi-Hierarchical Coherence Network, Global Semantics, Dialogue systems, Discourse Coherence

ACM Reference Format:
Jinhua Peng, Zongyang Ma, Di Jiang, Hua Wu. 2019. Integrating Bayesian and Neural Networks for Discourse Coherence. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316582

1 INTRODUCTION

Discourse coherence is a concept that measures semantic relevance between a natural language utterance and its context [14]. In the dialogue system, it usually retrieves or generates a set of responses based on a user's utterance. The function of discourse coherence is to rank and select the most appropriate one from the above candidate responses as the reply. This is crucial for the dialogue system, since inconsistent response might hurt the user's experience when interacting with a chatbot. In Figure 1, we present two cases to illustrate the utility of discourse coherence. In each case, the last utterance is a candidate reply of the dialogue system and all its previous utterances are the context. In the first case, *You almost sleep for 11 hours everyday!* is consistent with the theme of the context. On the contrary, in the second case, *Gigantic appetite* seems to be a dramatic shift of the dialogue theme and is not suitable for the given context. By qualitatively evaluating the semantic relevance between candidate replies and the context, discourse coherence enables dialogue systems to filter out inappropriate replies and pinpoint the appropriate ones.



		Chinese	English
Coherent Case	Context	A: … B: … A: 别睡了,快点起来开工了 B: 俺好累好困啊	A: … B: … A: Don' t sleep, get up to work! B: I am so tired and sleepy.
		A: 每天睡十一小时	A: You almost sleep for 11 hours everyday!
Incoherent Case	Context	A: … B: … A: 甜点好吃,使我无法自拔 B: 这个看着就好吃啊	A: … B: … A: The dessert tastes good, and I like it very much! B: It looks very delicious!
		A: 超级惊人的饭量	A: Gigantic appetite.

Figure 1: Real-life Dialogue Cases. The original Chinese texts are translated into English to enhance readability.

However, to the best of our knowledge, little work has been done to design specialized models for discourse coherence. Existing work [14] usually utilizes primitive textual information and conventional discriminative/generative models for evaluating discourse coherence, while ignores the topical interaction between utterances. Since these models are not initially proposed for discourse coherence, the capabilities of these methods are heavily limited in real-life applications. Considering the complicated semantic nature of dialogue data, we propose a novel framework that takes full advantage of Bayesian and neural networks for discourse coherence. The framework consists of two major components: the Bayesian component and the neural component, which are described as follows.

Table 1: Notations for cpLDA

z_i	the topic of sentence i
$\mathbf{z}_{-i}$	topics of all the sentences except that of sentence i
$\mathbf{v}_i$	set of words in sentence i
n_{dk}	number of words in document d assigned with topic k
n_{kv}	number of word v assigned with topic k
n_{di}	number of words in sentence i of document d
n_{iv}	number of word v in sentence i

The Bayesian component, which is named as cpLDA (short for Coherence-Pivoted Latent Dirichlet Allocation), encodes our prior

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316582

knowledge of dialogue data into topic-level semantics. cpLDA ensures both global and local semantic consistency for latent topic discovery. Hence, the resultant topics of cpLDA are influenced by the global word relations as well as the local structures within dialogue data (i.e., words in a sentence are constrained to one unique topic). cpLDA is well designed for mining latent topics from dialogue data, in which each utterance is relatively short and lacks of information. On the other hand, the neural component is a deep neural network model, naming MHCN (short for Multi-Hierarchical Coherence Network), equipped with the generated topics from cpLDA for distinguishing the coherent utterances from incoherent ones given the context. MHCN is essentially a three-layer Neural Network (i.e., word-level, sentence-level and combination-level layers), and takes various factors into account, including the Bag-of-Words representation and topic distribution of each utterance.

In order to verify the effectiveness of the proposed framework, we quantitatively compare it with the state-of-the-arts in terms of a variety of metrics. Extensive experimental results show that the proposed framework achieves superior performance and applicability in the dialogue system for the task of response selection.

The remainder of this paper is organized as follows. In Section 2, we discuss the related work. In Section 3, we discuss the technical details of our framework consisting of cpLDA and MHCN models. In Section 4, we report the experimental results of our proposed models for discourse coherence. Finally, we conclude the paper in Section 5.

2 RELATED WORK

In this section, we begin with a brief overview of the studies on discourse coherence. We then survey the related work about topic models.

The study of discourse coherence is motivated by many work about linguistic theories. Guinaudeau and Strub ([10]) extended the entity-grid model by converting the matrix to the bipartite graph, which is able to address computational complexity and data sparsity problems occurred in the entity-grid model. In [9], the authors combined the entity-grid model and the HMM-based content model in a unified framework and demonstrated that their proposed model outperforms entity-grid model on the task of sentence ordering(2007). Recently, Neural Net Models are proposed in [14] for the task of discourse coherence. The authors introduced two classes of neural models: discriminative models for distinguishing coherent discourse from incoherent one and generative models for producing coherent discourse. More specifically, the generative models are derived from the variational latent variable models, which treat the topic z of current utterance as a stochastic latent variable conditioned on all previous utterances. With the topic latent variable, these generative models perform best on the task of sentences reordering. Dissimilar to the study in [14], we explicitly incorporate the topical embedding into our MHCN model. That is, the topics are preliminarily discovered by cpLDA, which could be more flexible to add into the MHCN model.

The proposed cpLDA can be considered as a deeply restructured variant of LDA. LDA was firstly introduced in [4], where it was developed for discovering hidden topics from a large collection of documents. Various extensions are proposed due to the promising performance of LDA. In the following, we name a few. Dynamic Topic Model [3] incorporates the time attribute, assuming that topics are evolving with time. That is, the same topic can be represented by different sets of keywords during different time periods. Supervised Topic Model was proposed in [8], where topics are regularized by the response type (i.e., document label). In LDA, topics are regarded as independent factors. However, in common sense, topics are usually related with each other. To address this issue, Correlated Topic Model [2] models the relations between topics in a more natural way. Instead of assigning a topic to a word, in sentenceLDA [11], a unique topic is assigned to all the words in a sentence.

3 A FRAMEWORK INTEGRATING BAYESIAN AND NEURAL COMPONENTS

In this section, we present our framework, which consists of a Bayesian component (cpLDA) and a neural component (MHCN).

3.1 cpLDA

Here we discuss the technical details of cpLDA. In Section 3.1.1, we discuss how to impose the local semantics consistency by modeling the "sentences" in the generative process of cpLDA. In Section 3.1.2, we discuss how to achieve global semantics consistency by utilizing the knowledge about word relations.

3.1.1 Local Semantics Consistency. cpLDA associates a latent topic with each observation, which is the occurrence of a "sentence" and its "words" in a particular "document". We utilize d to denote a "document", w a "word", s a "sentence" and z a "latent topic". θ_d and ϕ_k represents the topic distribution of document d and the word distribution of topic k respectively. We also introduce the hyper-parameters α and β. Based on these notations, we present the generative process of cpLDA as follows:

- for each topic z_k, generate the word distribution $\phi_k \sim Dirichlet(\beta)$;
- for each document d, generate the topic distribution $\theta_d \sim Dirichlet(\alpha)$;
- for each sentence s in d:
  - draw a topic $z_k \sim Multinomial(\theta_d)$;
  - draw each word $w \sim Multinomial(\phi_k)$;

In the generative process above, we constrain that the words of a sentence are generated by the same topic, in order to ensure the local semantics consistency. Hence, cpLDA utilizes "sentences" rather than "words" as the basic units for topic assignment. For inferring cpLDA, we need to compute the posterior distribution of the hidden variables z, θ and ϕ as follows:

$$p(\theta, \phi, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \phi, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}$$

However, due to the coupling between θ and ϕ, this distribution is intractable to compute. Gibbs Sampling is an alternative algorithm for estimating the latent variables in cpLDA. The condition probability $p(z_i = k|\mathbf{z}_{-i}, \mathbf{w})$ is shown in Table 2.

The notations are explained in Table 1. Though Gibbs sampler is a simple method to infer cpLDA, it suffers from the high computational complexity, which limits its usage on large-scale data

Table 2: Conditional probability of topic z_i

$$p(z_i = k|\mathbf{z}_{-i}, \mathbf{w}) = \frac{n_{dk} + \alpha_k}{\sum_{k'=1}^{K}(n_{dk'} + \alpha_{k'})} \frac{\Gamma(\sum_{v'=1}^{V}(n_{kv'} + \beta_{v'}))}{\Gamma(\sum_{v'=1}^{V}(n_{kv'} + \beta_{v'}) + n_{di})} \prod_{v' \in \mathbf{v}_i} \frac{\Gamma(n_{kv'} + \beta_{v'} + n_{iv'})}{\Gamma(n_{kv'} + \beta_{v'})}$$

$$p'(z_i = k|\mathbf{z}_{-i}, \mathbf{w}) = \frac{n_{dk} + \alpha_k}{\sum_{k'=1}^{K}(n_{dk'} + \alpha_{k'})} \prod_{v \in \mathbf{v}_i} \frac{n_{kv} + \beta_v}{\sum_{v'=1}^{V}(n_{kv'} + \beta_{v'})}$$

set. In this paper, we introduce Metropolis-Hastings (MH) algorithm for cpLDA. Compared with Gibbs Sampling, when sampling a topic z, the time complexity of Gibbs Sampling is $O(K)$ (K represents the number of topics), while that of MH algorithm is approximately reduced to $O(1)$. The approximate conditional probability $p'(z_i = k|\mathbf{z}_{-i}, \mathbf{w})$ is shown in Table 2.

In order to proceed MH algorithm, several proposals are required. We define two proposals: Doc-Topic Proposal and Topic-Term Proposal. Doc-Topic Proposal is formulated as $\rho_d(k) \propto \frac{n_{dk}+\alpha_k}{\sum_{k'=1}^{K}(n_{dk'}+\alpha_{k'})}$ and the acceptance probability of topic transition from k to k' is $min\{1, \frac{p(k')\rho_d(k)}{p(k)\rho_d(k')}\}$. $p(k)$ can be selected according to the first conditional probability in Table 2. Similarly, the Topic-Term Proposal is formulated as $\rho_v(k) \propto \frac{n_{kv}+\beta_v}{\sum_{v'=1}^{V}(n_{kv'}+\beta_{v'})}$. We further create the alias table to accelerate the sampling process. When using the alias table, generating a topic z only requires two uniform samples, more details of alias table can be referred to [18].

3.1.2 *Global Semantics Consistency.*

The previous section illustrates how to model the "sentences" in cpLDA and details the approach of inferring the parameters via the MH algorithm. In this subsection, we discuss how to utilize global semantics consistency in cpLDA. The knowledge about word relations can be obtained from many sources such as human-engineering ontology [16] and automatically built knowledge bases [7]. In the scenario of dialogue system, high-quality word relations can be easily obtained by computing the similarity of word embedding vectors from an external knowledge base (e.g., microblog).

Assuming that the vocabulary size is W, the word relations are characterized by a $W \times W$ matrix $\mathbf{R}$. We denote the weight of the relation between two words w_a and w_b as R_{ab} and calculate it as follows:

$$R_{ab} = sim(v_a, v_b).$$

where v_a and v_b denote the word embedding vectors of w_a and w_b respectively.

We proceed to discuss the strategy of utilizing $\mathbf{R}$ in cpLDA. We want the probability $p(w|z_k)$ (i.e., ϕ_k^w) to be dependent on the word relation information stored in $\mathbf{R}$. Here we use a quadratic-form influence term with a trade-off factor τ. Formally, given $\mathbf{R}$, we adjust the topic-word distribution $p(w|z_k)$ as follows:

$$p(w|z_k) \leftarrow p(w|z_k) + \tau \frac{p(w|z_k) \sum_{i=1}^{W} R_{iw} p(i|z_k)}{P(\cdot|z_k)^T \mathbf{R} P(\cdot|z_k)}. \quad (1)$$

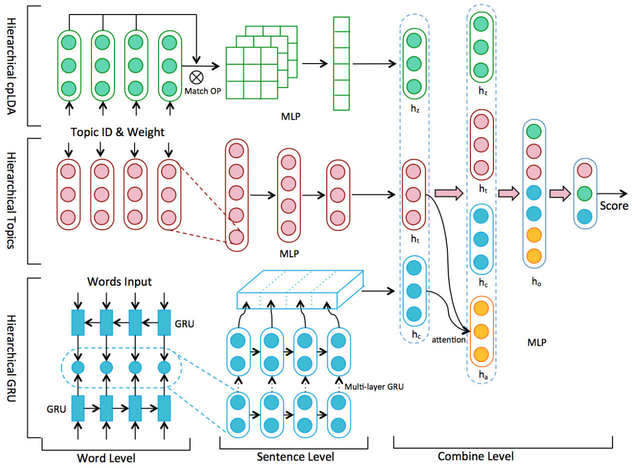


Figure 2: Multi-Hierarchical Coherence Network

It is easy to see that the adjusted $p(w|z_k)$ is influenced by the prominence of the other "words" that has relations with w in $\mathbf{R}$. In practice, Eq. (1) is applied after MH iteration until convergence is achieved. Since we are only interested in relatively frequent words from the vocabulary, $\mathbf{R}$ will be a sparse matrix and hence computations are efficient for $\mathbf{R}$ to be used in practice.

In summary, the global and local semantics consistency modeled in cpLDA fits the characteristics of dialogue data well. The utterance in the dialogue is one type of User Generated Content (UGC) data and each multiple-turn conversation could be treated as the "virtual document". Compared with the traditional web and news documents, this "virtual document" is much shorter with a few sentences, indicating that it contains relatively poor information. Therefore, in cpLDA, globally modeling word relations (global semantics coherence) is able to introduce the external knowledge and enrich the representation of the utterances. On the other hand, utterances are informal text, and full of abbreviations and slangs. That is, constraining each utterance with one unique topic (local semantics coherence) is beneficial for focusing on the key semantics of the utterance.

3.2 MHCN: A Coherence Network Model with Topic Embedding and Topic Attention

In this section, we describe our MHCN in details. MHCN is equipped with the GRU cells, and at the same time, it accepts the input of topic embeddings generated from the cpLDA model. Let $\mathbf{c}$ denote a sequence of coherent utterances in one session of the dialogue system. More specifically, $\mathbf{c} = \{u_1, u_2, ...u_L\}$, where u and L represents the utterance and the number of utterances respectively. u consists of a sequence of words $\{w_{u,1}, w_{u,2}, ...w_{u,n}\}$. Each word w is described by an embedding vector with the fixed dimensionality. Since the aim of this research is to calculate the coherence degree between the current utterance and all its previous contextual utterances in the same discourse, to make clear, we also use r (response) to denote u_L and $\mathbf{c}^{-r}$ to denote the contextual utterances. The data set is represented as the collection of triple $(\mathbf{c}_i^{-r}, r_i, y_i)$, where y_i is the label. $y_i = 1$ indicates that the response r_i is coherent with the contextual utterances $\mathbf{c}_i^{-r}$. Given the $(\mathbf{c}_i^{-r}, r_i)$ pair, MHCN model $F(\mathbf{c}_i^{-r}, r_i)$ attempts to calculate the coherence score y_i'.

The architecture of MHCN model is illustrated in Figure 2. MHCN is built with hierarchical layers (i.e., word-level, sentence-level and combination-level layers). MHCN accepts three types of input factors: topic ID, topic distribution of the utterance and word sequence of the utterance. (1) For the topic distribution, the similarity scores of the topic distributions between the response r and each utterance in $\mathbf{c}^{-r}$ are calculated. (2) The topic ID is inferred from the cpLDA. Observe that though cpLDA assumes that one utterance only occupies one topic, we concatenate the multiple sampling results for each utterance. In that case, each utterance has multiple topic IDs. The topic ID is treated as the "topical word" in our research and fed into the model as topical word embedding. (3) Each utterance is represented by a sequence of word embeddings, which are conventional embeddings and not pre-trained. Through the GRU cell (the word-level layer), the hidden representation of each utterance u is obtained. Then, these hidden utterance representations are propagated to another GRU cell (sentence-level layer). After this encoding process, the model yields the hidden representation of $\mathbf{c}$. Note that both of the GRU cells (i.e., word-level and sentence-level layers) are bidirectional. The third layer is the combination-level layer, we perform a attention mechanism to obtain utterance attention weight distributions over the topic. In this layer, all these three factor hidden representations and the topic attention weight are combined to calculate the last score.

In MHCN, we could obtain the hidden vector $h_{u,i}$ of utterance u_i by feeding a sequence of embedding word vectors $\{v_{u,1}, v_{u,2}, ...v_{u,n}\}$ into the GRU cell. $h_{u,i}$ can be calculated as follows:

$$
\begin{aligned}
a_i &= \sigma(\mathbf{W_a} v_{u,i} + \mathbf{U_a} h_{u,i-1}) \\
b_i &= \sigma(\mathbf{W_b} v_{u,i} + \mathbf{U_b} h_{u,i-1}) \\
h'_{u,i} &= tanh(\mathbf{W_h} v_{u,i} + \mathbf{U_a}(b_i \odot h_{u,i-1})) \\
h_{u,i} &= a_i \odot h'_{u,i} + (1 - a_i) \odot h_{u,i-1}
\end{aligned}
$$

where $\mathbf{W_a}$, $\mathbf{U_a}$, $\mathbf{W_b}$, $\mathbf{U_b}$, $\mathbf{W_h}$ and $\mathbf{U_a}$ are the parameters to be learned. a_i and b_i denote the update gate and reset gate respectively.

Similarly, the hidden vector h_c of $\mathbf{c}$ is obtained by feeding a sequence of hidden vector of utterances $h_{u_1}, h_{u_2}, ...h_{u_n}$. The advantage of using GRU cells is that it enables to keep and propagate key information as well as filter out noisy formation. Meanwhile, GRU cells are computed faster than LSTM cells. After obtaining the hidden vector h_t, h_c, we use them to calculate the topic attention weight h_a as follows:

$$
\begin{aligned}
h_{a_i} &= (\mathbf{W}_{a_i} h_t + b_{a_i}) \odot h_c \\
h_a &= (h_{a_1}, h_{a_2}, h_{a_3}, ..., h_{a_m})
\end{aligned}
$$

In the combination-level layer, the h_o is concatenated by the hidden vectors of h_c, h_z (topic-similarity vector), h_t (topic-id vector) and h_a (utterance-topic attention), then h_o is forwarded to the softmax as follows:

$$
I(\mathbf{c}^{-r}, r) = softmax(\mathbf{W}_o h_o + b_o)
$$

where $\mathbf{W}_o$ and b_o are the parameters. To learn all these parameters, we minimize the cross entropy as follows:

$$
-\sum (y_i \log I(c_i^{-r}, r_i) + (1 - y_i) \log(1 - I(c_i^{-r}, r_i)))
$$

The MHCN architecture well integrates the word-level and topic-level information of utterances. Moreover, the MHCN model regards the topical similarity between sentences and the similarity between sentence embeddings as well, which makes the model robust to capture the topical coherence between utterances.

4 EXPERIMENTS

In this section, we present the experimental results[1]. In Section 4.1, we evaluate the proposed framework in terms of Precision, Recall and F1. In Section 4.2, we demonstrate the impact of model parameters, followed by a case study in Section 4.3.

4.1 Data set and Baseline

We first briefly describe the data sets[2] used in our study. Then several the state-of-the-art baselines are introduced. Finally, we report the performance of cpLDA and MHCN for the task of discourse coherence.

Ubuntu Dialogue Corpus (UDC)[3] is derived from the Ubuntu Relay Chat Channel [17]. This dataset contains about 1.85 million conversations with an average of 5 utterances per conversation.

Weibo. To learn the cpLDA model, we introduce a Weibo data set, which consists of 1.1 million status-comment pairs. We treat each status-comment pair as a single-turn conversation and each status or comment as one sentence. We filter out the single term and stopwords in the data set. After preprocessing, the vocabulary size of the Weibo data set is 33,067.

Baidu Tieba. We construct our dialogue data set from Baidu Tieba, which serves as the largest Chinese forum and allows users to post status or comment on any topic. Our Tieba data consists of 1 million dialogues with multiple turns. We treat the current utterance as the response one and all the previous utterances in the same dialogue as the context. We regard the response in the dialogue as the positive one, and randomly sample the negative response from the Tieba data.

Baselines. We present two types of baselines: discriminative models and generative models.

- **Discriminative models** include *LSTM*, *CNN* [12], *Hierarchical RNN (HRNN)* [15], *MLP* and *HGRU* [15].

[1] In order to enhance reproducibility, the source code of the proposed framework is provided in the supplementary material.
[2] The data sets will be made publicly accessible upon publication.
[3] http://www.iulianserban.com/Files/UbuntuDialogueCorpus.zip

Table 3: Results of Discourse Coherence Evaluation

Methods	UDC			Tieba		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*
LSTM	0.7912	0.7961	0.7905	0.6230	0.6245	0.6229
CNN	0.8275	0.8263	0.8278	0.6253	0.6268	0.6250
HRNN	0.8317	0.8401	0.8309	0.5830	0.5841	0.5817
MLP	0.8304	0.8302	0.8273	0.6212	0.6222	0.6205
HGRU	0.8392	0.8415	0.8394	0.6421	0.6441	0.6404
SEQ2SEQ	0.7528	0.7591	0.7513	0.5530	0.5589	0.5517
SEQ2SEQ-MMI	0.7532	0.7410	0.7408	0.6023	0.6014	0.5996
VLV-GM	0.7986	0.7973	0.7869	0.6506	0.6503	0.6498
MHCN	**0.8969**	**0.8930**	**0.8938**	**0.7341**	**0.7319**	**0.7284**
MHCN(SLDA)	0.8691	0.8656	0.8634	0.6644	0.6652	0.6635
w/o H-cpLDA	0.8355	0.8368	0.8376	0.6533	0.6541	0.6516
w/o H-topic	0.8241	0.8244	0.8225	0.6482	0.6478	0.6459
w/o HGRU	0.7953	0.7917	0.7933	0.6129	0.6105	0.6146
w/o Attention	0.8740	0.8719	0.8682	0.7103	0.7114	0.7062

- **Generative models** employed in our study are *SEQ2SEQ* [1], *SEQ2SEQ-MMI* [14] and *VLV-GM* [14]. *SEQ2SEQ* is an encoder-decoder model with the attention mechanism to predict the current utterance given its context. *SEQ2SEQ-MMI* can be regarded as the advanced *SEQ2SEQ* model with the objective function Maximum Mutual Information (MMI). As reported in [14], *SEQ2SEQ-MMI* yields more diverse responses compared with *SEQ2SEQ*. *VLV-GM* is a model that combines encoder-decoder and a Markov chain using global information z_n like LDA. *VLV-GM* is closely related to many attempts in training variational autoencoders(VAE) [13], variational or latent-variable recurrent nets [5, 6], hierarchical latent variable encoder-decoder models [17].

Experimental Setup. For all models, we set the dimensionality of word vectors to 200. We adopt the stochastic gradient descent (SGD) and the learning rate is set to 0.0005. The batch size is 1024 and the epoch times is 200. All the models are implemented using tensorflow and trained on a Tesla P40 GPU. For the cpLDA and sentenceLDA, we experimentally set the topic number to 200 and the epoch times to 400. The cpLDA and sentenceLDA are learned from the Weibo data set and the UDC data, and then applied to infer the topic of utterance for the Tieba data set and the UDC data.

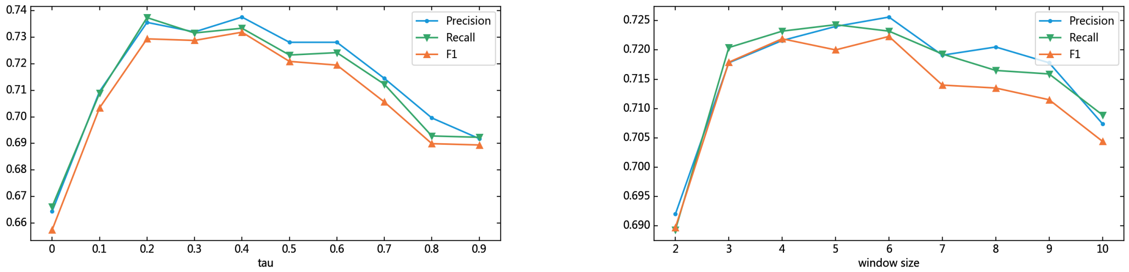
4.2 Results Analysis and Parameter Sensitivity

Results Analysis. Table 3 demonstrates the experimental results. We could observe that our MHCN model outperforms all state-of-the-art models measured in terms of all metrics. In order to investigate the influence of different components of our model, we also replace cpLDA with SLDA which is short for sentenceLDA. In specific, we discard hierarchical cpLDA (noted as w/o H-cpLDA), hierarchical topic embedding (noted as w/o H-topic), hierarchical GRU (noted as w/o HGRU) and attention mechanism in combination-layer (noted as w/o Attention), respectively.

We summarize the observation as follows: (1) *SEQ2SEQ* yields the worst results measured by precision, indicating that the generative models could not well distinguish positive responses from negative ones. One possible reason is that the generative models are more likely to yield safe or simple response, which can not match the real response well. (2) *MHCN* achieves better performance than both *HGRU* and *MHCN(SLDA)*. It proves that the mined topic embeddings facilitate the performance of the *MHCN* model. Moreover, Compared with *MHCN(SLDA)*, *MHCN* achieves significant improvements in terms of all metrics, which justifies the capability of cpLDA on generating fine-grained topics. (3) In *MHCN* achieves, each parts make contributions to the final performance. In specific, the attention mechanism find out the common information of topic representations and utterance representations. Therefore it makes topic and utterance representations combine more closely.

Parameter Sensitivity. We proceed to investigate the influence of model parameters. First, we tune the trade-off parameter τ for global coherence in cpLDA (see Eq. 1). Recall that τ is introduced to balance the local and global coherences of topics, and a larger τ indicates that the cpLDA is more influenced by the global coherence. We vary τ from 0.0 to 0.9 with the step 0.1 and train the cpLDA models, then the resultant topics inferred from the corresponding cpLDA are fed to the MHCN model to measure the performance of discourse coherence. The experimental results are reported in Figure 3(a). We observe that τ slightly influences the merits (i.e., precision, recall and F1) when τ varies from 0.1 to 0.9. In addition, the MHCN model with $\tau \in [0.2, 0.4]$ achieves the best performance. However, when τ is set to 0.0, meaning that only local coherence is considered (same to the sentenceLDA model), the performance measured by any merit drops sharply. This observation shows the importance of the global coherence in cpLDA.

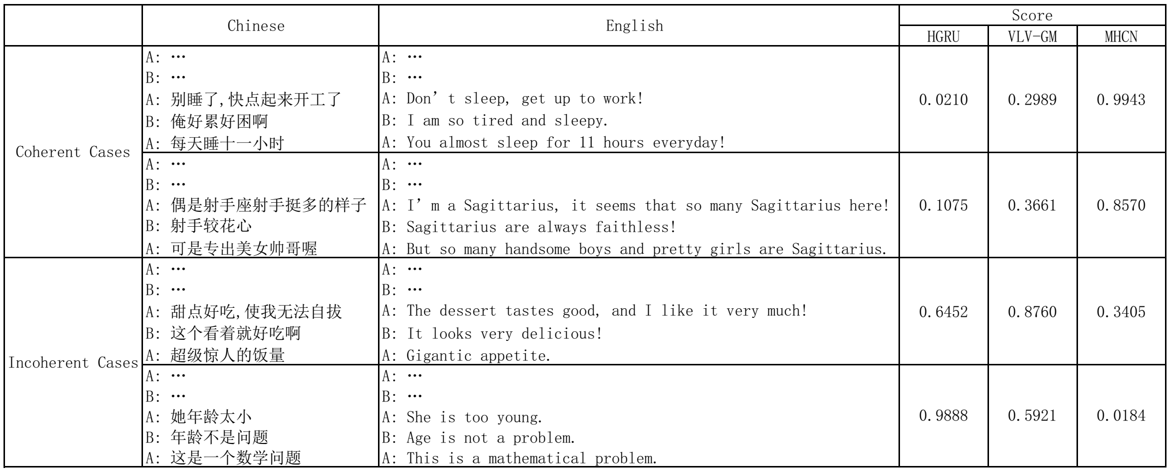
Next, we investigate the parameter window size W on the model performance. W represents the number of utterances in each discourse. Recall that in Section 4.1 we set the window size to the fixed number 6. Here we slide the window on the discourse and select number of utterances ranging from 2 to 10. In another word, we evaluate the coherence between $W - 1$ utterances and the current utterance. The experimental results are shown in Figure 3(b). We discover that the MHCN model with $W \in (3, 4, 5, 6)$ achieves the best performance, and then slightly declines.



(a) Varying Trade-off Parameter τ

(b) Varying Window Size W

Figure 3: Parameter Sensitivity



	Chinese	English	Score		
			HGRU	VLV-GM	MHCN
Coherent Cases	A：… B：… A：别睡了,快点起来开工了 B：俺好累好困啊 A：每天睡十一小时	A：… B：… A：Don’ t sleep， get up to work! B：I am so tired and sleepy. A：You almost sleep for 11 hours everyday!	0.0210	0.2989	0.9943
	A：… B：… A：偶是射手座射手挺多的样子 B：射手较花心 A：可是专出美女帅哥喔	A：… B：… A：I’ m a Sagittarius， it seems that so many Sagittarius here! B：Sagittarius are always faithless! A：But so many handsome boys and pretty girls are Sagittarius.	0.1075	0.3661	0.8570
Incoherent Cases	A：… B：… A：甜点好吃,使我无法自拔 B：这个看着就好吃啊 A：超级惊人的饭量	A：… B：… A：The dessert tastes good， and I like it very much! B：It looks very delicious! A：Gigantic appetite.	0.6452	0.8760	0.3405
	A：… B：… A：她年龄太小 B：年龄不是问题 A：这是一个数学问题	A：… B：… A：She is too young. B：Age is not a problem. A：This is a mathematical problem.	0.9888	0.5921	0.0184

Figure 4: Case Study

Finally, the phenomenon indicates that τ and W as the value increases, the performance first increases and then decreases. One possible reason is that the more information is provided, the more noise is introduced.

4.3 Case Study

We show several cases in Figure 4. Each case contains 3 utterances (one current utterance and two contextual utterances) and coherence scores generated by different methods. The coherence score is calculated as the coherence degree between the current utterance and its context. We quantitatively compare our *MHCN* with the best discriminative baseline *HGRU* and the best generative baseline *VLV-GM*. For the first case on topic “sleep", the current utterance “You almost sleep for 11 hours everyday!" is topically coherent with its two previous utterances (i.e., “Don’t sleep, get up to work!" and “I am so tired and sleepy."). While *VLV-GM* and *HGRU* yield lower scores 0.2989 and 0.0210. In contrast, *MHCN* obtains much larger coherence score 0.9943. Similar observations can been seen in the second case on topic “constellation". For the third case, the current utterance “gigantic appetite." is obviously irrelevant with its two previous utterances on topic “dessert". Only *MHCN* detects the incoherence between utterances and assigns the smallest score. Similar result is observed for the last case. Due to the page limination, we only show these four cases. In practice, we analyze a massive amount of cases, and find that, compared with the state-of-the-art discriminative and generative baselines, our neural structure with the fed topic embedding is able to capture the semantic relation between utterances more precisely. That is, the topic information of utterances is well encoded and decoded in *MHCN*, facilitating the computation of the coherence degree between utterances. Our *MHCN* model is superior to other state-of-the-art models in the discourse coherence task for the dialogue system.

5 CONCLUSION

In this paper, we tackle with the problem of discourse coherence. The main objective of this research is to distinguish the coherent utterances from incoherent ones given the context. To address the problem, we propose a novel framework that seamlessly integrates Bayesian and neural components. These two components are Coherence-Pivoted Latent Dirichlet Allocation (cpLDA) and

Multi-Hierarchical Coherence Network (MHCN). The former aims to infer fine-grained topics by considering both local and global semantics coherence. The latter employs multiple GRU cells and takes full advantage of the topics generated by cpLDA. Experimental results show that the proposed framework achieves superior performance comparing to several state-of-the-art methods.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *Computer Science* (2014).

[2] David M. Blei and John D. Lafferty. 2005. Correlated Topic Models. In *Advances in Neural Information Processing Systems 18*. 147–154.

[3] David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Machine Learning, Proceedings of the Twenty-Third International Conference*. 113–120.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[5] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating Sentences from a Continuous Space. *Computer Science* (2015).

[6] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron Courville, and Yoshua Bengio. 2015. A Recurrent Latent Variable Model for Sequential Data. *Computer Science* 35, 8 (2015), 1340–1353.

[7] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*. 601–610.

[8] Miroslav Dudík, David M. Blei, and Robert E. Schapire. 2007. Hierarchical maximum entropy density estimation. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference*. 249–256.

[9] Micha Elsner, Joseph L. Austerweil, and Eugene Charniak. 2007. A Unified Local and Global Model for Discourse Coherence. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*. 436–443.

[10] Camille Guinaudeau and Michael Strube. 2013. Graph-based Local Coherence Modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*. 93–103.

[11] Yohan Jo and Alice H. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining*. 815–824.

[12] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A Convolutional Neural Network for Modelling Sentences. *Eprint Arxiv* 1 (2014).

[13] Diederik P Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. (2013).

[14] Jiwei Li and Dan Jurafsky. 2017. Neural Net Models of Open-domain Discourse Coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 198–209.

[15] Zhao Meng, Lili Mou, and Zhi Jin. 2017. Hierarchical RNN with Static Sentence-Level Attention for Text-Based Speaker Change Detection. (2017).

[16] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[17] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. (2016).

[18] Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Po Xing, Tie-Yan Liu, and Wei-Ying Ma. 2015. LightLDA: Big Topic Models on Modest Computer Clusters. In *Proceedings of the 24th International Conference on World Wide Web*. 1351–1361.