Hidden Softmax Sequence Model for Dialogue Structure Analysis

Zhiyang He[1]**, Xien Liu**[2]**, Ping Lv**[2]**, Ji Wu**[1]
[1]Department of Electronic Engineering, Tsinghua University, Beijing, China
[2]Tsinghua-iFlytek Joint Laboratory for Speech Technology, Beijing, China
{zyhe_ts, xeliu, luping_ts, wuji_ee}@mail.tsinghua.edu.cn

Abstract

We propose a new unsupervised learning model, hidden softmax sequence model (HSSM), based on Boltzmann machine for dialogue structure analysis. The model employs three types of units in the hidden layer to discovery dialogue latent structures: softmax units which represent latent states of utterances; binary units which represent latent topics specified by dialogues; and a binary unit that represents the global general topic shared across the whole dialogue corpus. In addition, the model contains extra connections between adjacent hidden softmax units to formulate the dependency between latent states. Two different kinds of real world dialogue corpora, Twitter-Post and AirTicketBooking, are utilized for extensive comparing experiments, and the results illustrate that the proposed model outperforms sate-of-the-art popular approaches.

1 Introduction

Dialogue structure analysis is an important and fundamental task in the natural language processing domain. The technology provides essential clues for solving real-world problems, such as producing dialogue summaries (Murray et al., 2006; Liu et al., 2010), controlling conversational agents (Wilks, 2006), and designing interactive dialogue systems (Young, 2006; Allen et al., 2007) etc. The study of modeling dialogues always assumes that for each dialogue there exists an unique latent structure (namely dialogue structure), which consists of a series of latent states.[1]

Some past works mainly rely on supervised or semi-supervised learning, which always involve extensive human efforts to manually construct latent state inventory and to label training samples. Cohen et al. (2004) developed an inventory of latent states specific to E-mail in an office domain by inspecting a large corpus of e-mail. Jeong et al. (2009) employed semi-supervised learning to transfer latent states from labeled speech corpora to the Internet media and e-mail. Involving extensive human efforts constrains scaling the training sample size (which is essential to supervised learning) and application domains.

In recent years, there has been some work on modeling dialogues with unsupervised learning methods which operate only on unlabeled observed data. Crook et al. (2009) employed Dirichlet process mixture clustering models to recognize latent states for each utterance in dialogues from a travel-planning domain, but they do not inspect dialogues' sequential structure. Chotimongkol (2008) proposed a hidden Markov model (HMM) based dialogue analysis model to study structures of task-oriented conversations from in-domain dialogue corpus. More recently, Ritter et al. (2010) extended the HMM based conversation model by introducing additional word sources for topic learning process. Zhai et al. (2014) assumed words in an utterance are emitted from topic models under HMM framework, and topics were shared across all latent states. All these dialogue structure analysis models are directed generative models, in which the HMMs, language models and topic models are combined together.

In this study, we attempt to develop a Boltzmann machine based **undirected generative model** for dialogue structure analysis. As for the document modeling using undirected generative model, Hinton and Salakhutdinov (2009) proposed a general framework, replicated soft-

[1] Also called dialogue acts or speech acts in some past work. In this paper, for simplicity we will only use the term "latent state" to describe the sequential dialogue structure.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 2063–2072,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

max model (RSM), for topic modeling based on restricted Boltzmann machine (RBM). The model focuses on the document-level topic analysis, it cannot be applied for the structure analysis. We propose a hidden softmax sequence model (HSSM) for the dialogue modeling and structure analysis. HSSM is a two-layer special Boltzmann machine. The visible layer contains softmax units used to model words in a dialogue, which are the same with the visible layer in RSM (Hinton and Salakhutdinov, 2009). However, the hidden layer has completely different design. There are three kinds of hidden units: softmax hidden units, which is utilized for representing latent states of dialogues; binary units used for representing dialogue specific topics; and a special binary unit used for representing the general topic of the dialogue corpus. Moreover, unlike RSM whose hidden binary units are conditionally independent when visible units are given, HSSM has extra connections utilized to formulate the dependency between adjacent softmax units in the hidden layer. The connections are the latent states of two adjacent utterances. Therefore, HSSM can be considered as a special Boltzmann machine.

The remainder of this paper is organized as follows. Section 2 introduces two real world dialogue corpora utilized in our experiments. Section 3 describes the proposed hidden softmax sequence model. Experimental results and discussions are presented in Section 4. Finally, Section 5 presents our conclusions.

2 Data Set

Two different datasets are utilized to test the effectiveness of our proposed model: a corpus of post conversations drawn from Twitter (Twitter-Post), and a corpus of task-oriented human-human dialogues in the airline ticket booking domain (AirTicketBooking).

2.1 Twitter-Post

Conversations in Twitter are carried out by replying or responding to specific posts with short 140-character messages. The post length restriction makes Twitter keep more chat-like interactions than blog posts. The style of writing used on Twitter is widely varied, highly ungrammatical, and often with spelling errors. For example, the terms "be4", "b4", and "bef4" are always appeared in the Twitter posts to represent the word "before". Here, we totally collected about 900, 000 raw Twitter dialogue sessions. The majority of conversation sessions are very short; and the frequencies of conversation session lengths follow a power law relationship as described in (Ritter et al., 2010). For simplicity , in the data preprocessing stage non-English sentences were dropped; and non-English characters, punctuation marks, and some non-meaning tokens (such as "&") were also filtered from dialogues. We filtered short Twitter dialogue sessions and randomly sampled 5,000 dialogues (the numbers of utterances in dialogues rang from 5 to 25) to build the Twitter-Post dataset.

2.2 AirTicketBooking

The AirTicketBooking corpus consists of a set of task-oriented human-human mandarin dialogues from an airline ticket booking service center. The manual transcripts of the speech dialogues are utilized in our experiments. In the dataset, there is always a relative clear structure underlying each dialogue. A dialogue often begins with a customer's request about airline ticket issues. And the service agent always firstly checks the client's personal information, such as name, phone number and credit card numberm, etc. Then the agent starts to deal with the client's request. We totally collected 1,890 text-based dialogue sessions obtaining about 40,000 conversation utterances with length ranging from 15 to 100.

3 Dialogue Structure Analysis

3.1 Model Design

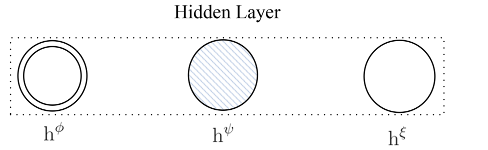

Figure 1: Hidden layer that consists of different types of latent variables

We design an undirected generative model based on Boltzmann machine. As we known, dialogue structure analysis models are always based on an underlying assumption: each utterance in the dialogues is generated from one latent state, which has a causal effect on the words. For instance, an utterance in AirTicketBooking dataset, "Tomorrow afternoon, about 3 o'clock" corresponds to the latent state "Time Information". However, by carefully examining words in dialogues we can observe that not all words are generated from the latent states (Ritter et al., 2010; Zhai and Williams, 2014). There are some words relevant to a global or background topic shared across dialogues. For example, "about" and "that" belong to a global (general English) topic. Some other words in a dialogue may be strongly related to the dialogue specific topic. For example, "cake", "toast" and "pizza" may appear in a Twitter dialogue with respect to a specific topic, "food". From the perspective of generative model, we can also consider that words in a dialogue are generated by the mixture model of latent states, a global/background topic, and a dialogue specific topic. Therefore, there are three kinds of units in the hidden layer of our proposed model, which are displayed in Figure 1. $\mathbf{h}^{\phi}$ is a softmax unit, which indicates the latent state for a utterance. $\mathbf{h}^{\psi}$ and $\mathbf{h}^{\xi}$ represent the general topic, and the dialogue specific topic, respectively. For the visible layer, we utilize the softmax units to model words in each utterance, which is the same with the approach in RSM (Hinton and Salakhutdinov, 2009). In Section 3.2, We propose a basic model based on Boltzmann machine to formulate each word in utterances of dialogues.

A dialogue can be abstractly viewed as a sequence of latent states in a certain reasonable order. Therefore, formulating the dependency between latent states is another import issue for dialogue structure analysis. In our model, we assume that each utterance's latent state is dependent on its two neighbours. So there exist connections between each pair of adjacent hidden softmax units in the hidden layer. The details of the model will be presented in Section 3.3.

3.2 HSM: Hidden Softmax Model

Notation	Explanation
K	dictionary size
J	number of latent states
$\mathbf{V}$	observed visibles representing words in dialogues
$\mathbf{b}$	bias term of $\mathbf{V}$
$\mathbf{h}^{\phi}$	latent variables representing latent states
$\mathbf{h}^{\psi}$	latent variable representing corpus general topic
$\mathbf{h}^{\xi}$	latent variables representing dialogue specific topics
$\mathbf{a}^{\phi}$	bias terms of $\mathbf{h}^{\phi}$
$\mathbf{a}^{\psi}$	bias term of $\mathbf{h}^{\psi}$
$\mathbf{a}^{\xi}$	bias terms of $\mathbf{h}^{\xi}$
$\mathbf{W}^{\phi}$	weights connecting $\mathbf{h}^{\phi}$ to $\mathbf{V}$
$\mathbf{W}^{\psi}$	weights connecting $\mathbf{h}^{\psi}$ to $\mathbf{V}$
$\mathbf{W}^{\xi}$	weights connecting $\mathbf{h}^{\xi}$ to $\mathbf{V}$
$\mathbf{F}, \mathbf{F}^{s}, \mathbf{F}^{e}$	weights between hidden softmax units

Table 1: Definition of notations.

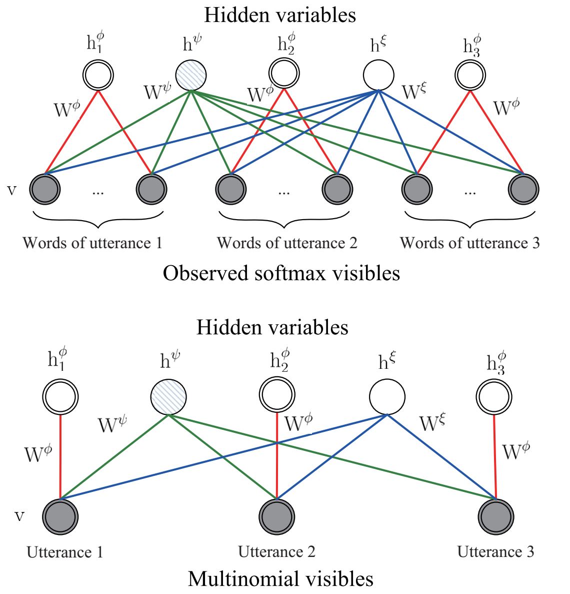

Figure 2: Hidden Softmax Model. The bottom layer are softmax visible units and the top layer consists of three types of hidden units: softmax hidden units used for representing latent states, a binary stochastic hidden unit used for representing the dialogue specific topic, and a special binary stochastic hidden unit used for representing corpus general topic. **Upper**: The model for a dialogue session containing three utterances. Connection lines in the same color related to a latent state represent the same weight matrix. **Lower**: A different interpretation of the Hidden Softmax Model, in which D_r visible softmax units in the r^{th} utterance are replaced by one single multinomial unit which is sampled D_r times.

Table 1 summarizes important notations utilized in this paper. Before introducing the ultimate learning model for dialogue structure analysis, we firstly discuss a simplified version, Hidden Softmax Model (HSM), which is based on Boltzmann machine and assumes that the latent variables are independent given visible units. HSM has a two-layer architecture as shown in Figure 2. The energy of the state $\{\mathbf{V}, \mathbf{h}^{\phi}, \mathbf{h}^{\psi}, \mathbf{h}^{\xi}\}$ is defined as follows:

$$\begin{aligned} E(\mathbf{V}, \mathbf{h}^{\phi}, \mathbf{h}^{\psi}, \mathbf{h}^{\xi}) =& \bar{E}_{\phi}(\mathbf{V}, \mathbf{h}^{\phi}) + \bar{E}_{\psi}(\mathbf{V}, \mathbf{h}^{\psi}) \\ &+ \bar{E}_{\xi}(\mathbf{V}, \mathbf{h}^{\xi}) + C(\mathbf{V}), \end{aligned} \tag{1}$$

where $\bar{E}_{\phi}(\mathbf{V}, \mathbf{h}^{\phi})$, $\bar{E}_{\psi}(\mathbf{V}, \mathbf{h}^{\psi})$ and $\bar{E}_{\xi}(\mathbf{V}, \mathbf{h}^{\xi})$ are sub-energy functions related to hidden variables $\mathbf{h}^{\phi}$, $\mathbf{h}^{\psi}$, and $\mathbf{h}^{\xi}$, respectively. $C(\mathbf{V})$ is the shared visible units bias term. Suppose K is the dictionary size, D_r is the r^{th} utterance size (i.e. the number of words in the r^{th} utterance), and R is the number of utterances in the a dialogue.

For each utterance $v_r(r = 1, .., R)$ in the dialogue session we have a hidden variable vector h_r^{ϕ} (with size of J) as a latent state of the utterance, the sub-energy function $\bar{E}_{\phi}(\mathbf{V}, \mathbf{h}^{\phi})$ is defined by

$$\begin{aligned}\bar{E}_{\phi}(\mathbf{V}, \mathbf{h}^{\phi}) = &-\sum_{r=1}^{R}\sum_{j=1}^{J}\sum_{i=1}^{D_r}\sum_{k=1}^{K} h_{rj}^{\phi} W_{rjik}^{\phi} v_{rik} \\ &-\sum_{r=1}^{R}\sum_{j=1}^{J} h_{rj}^{\phi} a_{rj}^{\phi},\end{aligned} \tag{2}$$

where $v_{rik} = 1$ means the i^{th} visible unit v_{ri} in the r^{th} utterance takes on k^{th} value, $h_{rj}^{\phi} = 1$ means the r^{th} softmax hidden units takes on j^{th} value, and a_{rj}^{ϕ} is the corresponding bias. W_{rjik}^{ϕ} is a symmetric interaction term between visible unit v_{ri} that takes on k^{th} value and hidden variable h_r^{ϕ} that takes on j^{th} value.

The sub-energy function $\bar{E}_{\psi}(\mathbf{V}, \mathbf{h}^{\psi})$, related to the global general topic of the corpus, is defined by

$$\bar{E}_{\psi}(\mathbf{V}, \mathbf{h}^{\psi}) = -\sum_{r=1}^{R}\sum_{i=1}^{D_r}\sum_{k=1}^{K} h^{\psi} W_{rik}^{\psi} v_{rik} - h^{\psi} a^{\psi}. \tag{3}$$

The sub-energy function $\bar{E}_{\xi}(\mathbf{V}, \mathbf{h}^{\xi})$ corresponds to the dialogue specific topic, and is defined by

$$\bar{E}_{\xi}(\mathbf{V}, \mathbf{h}^{\xi}) = -\sum_{r=1}^{R}\sum_{i=1}^{D_r}\sum_{k=1}^{K} h^{\xi} W_{rik}^{\xi} v_{rik} - h^{\xi} a^{\xi}. \tag{4}$$

W_{rik}^{ψ} in Eq. (3) and W_{rik}^{ξ} in Eq. (4) are two symmetric interaction terms between visible units and the corresponding hidden units, which are similar to W_{rjik}^{ϕ} in (2); a^{ψ} and a^{ξ} are the corresponding biases. $C(\mathbf{V})$ is defined by

$$C(\mathbf{V}) = -\sum_{r=1}^{R}\sum_{i=1}^{D_r}\sum_{k=1}^{K} v_{rik} b_{rik}, \tag{5}$$

where b_{rik} is the corresponding bias.

The probability that the model assigns to a visible binary matrix $\mathbf{V} = \{v_1, v_2, ..., v_D\}$ (where $D = \sum_{r=1}^{R} D_r$ is the dialogue session size) is

$$\begin{aligned}P(\mathbf{V}) &= \frac{1}{\mathcal{Z}} \sum_{\mathbf{h}^{\phi},\, \mathbf{h}^{\psi}, \mathbf{h}^{\xi}} \exp(-E(\mathbf{V}, \mathbf{h}^{\phi},\ \mathbf{h}^{\psi}, \mathbf{h}^{\xi})) \\ \mathcal{Z} &= \sum_{\mathbf{V}} \sum_{\mathbf{h}^{\phi},\, \mathbf{h}^{\psi}, \mathbf{h}^{\xi}} \exp(-E(\mathbf{V}, \mathbf{h}^{\phi},\ \mathbf{h}^{\psi},\ \mathbf{h}^{\xi}),\end{aligned} \tag{6}$$

where $\mathcal{Z}$ is known as the partition function or normalizing constant.

In our proposed model, for each word in the document we use a softmax unit to represent it. For the sake of simplicity, assume that the order of words in an utterance is ignored. Therefore, all of these softmax units can share the same set of weights that connect them to hidden units, thus the visible bias term $C(\mathbf{V})$ and the sub-energy functions $\bar{E}_{\phi}(\mathbf{V}, \mathbf{h}^{\phi})$, $\bar{E}_{\psi}(\mathbf{V}, \mathbf{h}^{\psi})$ and $\bar{E}_{\xi}(\mathbf{V}, \mathbf{h}^{\xi})$ in Eq. (1) can be redefined as follows:

$$\begin{aligned}\bar{E}_{\phi}(\mathbf{V}, \mathbf{h}^{\phi}) = &-\sum_{r=1}^{R}\sum_{j=1}^{J}\sum_{k=1}^{K} h_{rj}^{\phi} W_{jk}^{\phi} \hat{v}_{rk} \\ &-\sum_{r=1}^{R}(D_r \sum_{j=1}^{J} h_{rj}^{\phi} a_j^{\phi})\end{aligned} \tag{7}$$

$$\bar{E}_{\psi}(\mathbf{V}, \mathbf{h}^{\psi}) = -\sum_{k=1}^{K} h^{\psi} W_k^{\psi} \hat{v}_k - D h^{\psi} a^{\psi} \tag{8}$$

$$\bar{E}_{\xi}(\mathbf{V}, \mathbf{h}^{\xi}) = -\sum_{k=1}^{K} h^{\xi} W_k^{\xi} \hat{v}_k - D h^{\xi} a^{\xi} \tag{9}$$

$$C(\mathbf{V}) = -\sum_{k=1}^{K} \hat{v}_k b_k, \tag{10}$$

where $\hat{v}_{rk} = \sum_{i=1}^{D_r} v_{rik}$ denotes the count for the k^{th} word in the r^{th} utterance of the dialogue, $\hat{v}_k = \sum_{r=1}^{R} \hat{v}_{rk}$ is the count for the k^{th} word in whole dialogue session. D_r and D ($D = \sum_{r=1}^{R} D_r$) are employed as the scaling parameters, which can make hidden units behave sensibly when dealing with dialogues of different lengths (Hinton and Salakhutdinov, 2009).

The conditional distributions are given by softmax and logistic functions:

$$P(h_{rj}^{\phi} = 1|\mathbf{V}) = \frac{exp(\sum_{k=1}^{K} W_{jk}^{\phi} \hat{v}_{rk} + D_r a_j^{\phi})}{\sum_{j'=1}^{J} exp(\sum_{k=1}^{K} W_{j'k}^{\phi} \hat{v}_{rk} + D_r a_{j'}^{\phi})} \tag{11}$$

$$P(h^{\psi} = 1|\mathbf{V}) = \sigma(\sum_{k=1}^{K} W_k^{\psi} \hat{v}_k + D a^{\psi}) \tag{12}$$

$$P(h^{\xi} = 1|\mathbf{V}) = \sigma(\sum_{k=1}^{K} W_k^{\xi} \hat{v}_k + D a^{\xi}) \tag{13}$$

$$\begin{aligned}&P(v_{rik} = 1|\mathbf{h}^{\phi}, \mathbf{h}^{\psi}, \mathbf{h}^{\xi}) = \\ &\frac{exp(\sum_{j=1}^{J} h_{rj}^{\phi} W_{jk}^{\phi} + h^{\psi} W_k^{\psi} + h^{\xi} W_k^{\xi} + b_k)}{\sum_{k'=1}^{K} exp(\sum_{j=1}^{J} h_{rj}^{\phi} W_{jk'}^{\phi} + h^{\psi} W_{k'}^{\psi} + h^{\xi} W_{k'}^{\xi} + b_{k'})},\end{aligned} \tag{14}$$

where $\sigma(x) = 1/(1 + exp(-x))$ is the logistic function.

3.3 HSSM: Hidden Softmax Sequence Model

In this section, we consider the dependency between the adjacent latent states of utterances, and extend the HSM to hidden softmax sequence model (HSSM), which is displayed in Figure 3. We define the energy of the state $\{\mathbf{V}, \mathbf{h}^{\phi}, \mathbf{h}^{\psi}, \mathbf{h}^{\xi}\}$ in HSSM as follows:

$$\begin{aligned} E(\mathbf{V}, \mathbf{h}^{\phi}, \mathbf{h}^{\psi}, \mathbf{h}^{\xi}) =& \bar{E}_{\phi}(\mathbf{V}, \mathbf{h}^{\phi}) + \bar{E}_{\psi}(\mathbf{V}, \mathbf{h}^{\psi}) + \bar{E}_{\xi}(\mathbf{V}, \mathbf{h}^{\xi}) \\ &+ C(\mathbf{V}) + \bar{E}_{\Phi}(\mathbf{h}^{\phi}, \mathbf{h}^{\phi}), \end{aligned} \tag{15}$$

where $C(\mathbf{V})$, $\bar{E}_{\phi}(\mathbf{V}, \mathbf{h}^{\phi})$, $\bar{E}_{\psi}(\mathbf{V}, \mathbf{h}^{\psi})$ and $\bar{E}_{\xi}(\mathbf{V}, \mathbf{h}^{\xi})$ are the same with that in HSM. The last term $\bar{E}_{\Phi}(\mathbf{h}^{\phi}, \mathbf{h}^{\phi})$ is utilized to formulate the dependency between latent variables $\mathbf{h}^{\phi}$, which is defined as follows:

$$\begin{aligned} \bar{E}_{\Phi}(\mathbf{h}^{\phi}, \mathbf{h}^{\phi}) = &- \sum_{q=1}^{J} h_s^{\phi} F_q^s h_{1q}^{\phi} - \sum_{q=1}^{J} h_{Rq}^{\phi} F_q^e h_e^{\phi} \\ &- \sum_{r=1}^{R-1} \sum_{j=1}^{J} \sum_{q=1}^{J} h_{rj}^{\phi} F_{jq} h_{r+1,q}^{\phi}, \end{aligned} \tag{16}$$

where h_s^{ϕ} and h_e^{ϕ} are two constant scalar variables ($h_s^{\phi} \equiv 1$, $h_e^{\phi} \equiv 1$), which represent the virtual beginning state unit and ending state unit of a dialogue. F^s is a vector with size J, and its elements measure the dependency between h_s^{ϕ} and the latent softmax units of the first utterance. F^e also contains J elements, and in contrast to F^s, F^e represents the dependency measure between h_e^{ϕ} and the latent softmax units of the last utterance. F is a symmetric matrix for formulating dependency between each two adjacent hidden units pair $(h_r^{\phi}, h_{r+1}^{\phi})$, $r = 1, ..., R-1$.

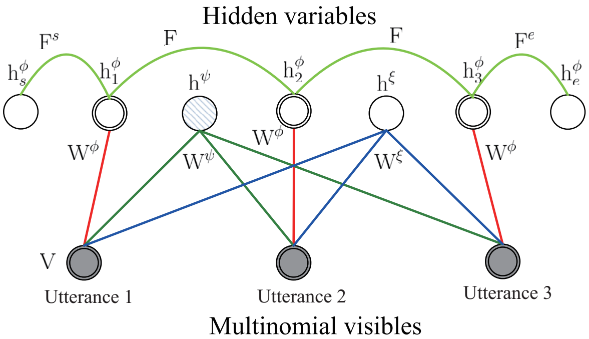

Figure 3: Hidden softmax sequence model. A connection between each pair of adjacent hidden softmax units is added to formulate the dependency between the two corresponding latent states.

3.4 Parameter Learning

Exact maximum likelihood learning in the proposed model is intractable. "Contrastive Divergence" (Hinton, 2002) can be used for HSM's learning, however, it can not be utilized for HSSM, because the hidden-to-hidden interaction term, $\{F, F^s, F^e\}$, result in the intractability when obtaining exact samples from the conditional distribution $P(h_{rj}^{\phi} = 1|\mathbf{V})$, $r = [1, R]$, $j \in [1, J]$. We use the mean-field variational inference (Hinton and Zemel, 1994; Neal and Hinton, 1998; Jordan et al., 1999) and a stochastic approximation procedure (SAP) (Tieleman, 2008) to estimate HSSM's parameters. The variational learning is utilized to get the data-dependent expectations, and SAP is utilized to estimate the model's expectation. The log-likelihood of the HSSM has the following variational lower bound:

$$\log P(\mathbf{V}; \theta) \geq \sum_{\mathbf{h}} Q(\mathbf{h}) \log P(\mathbf{V}, \mathbf{h}; \theta) + H(Q). \tag{17}$$

$Q(\mathbf{h})$ can be any distribution of $\mathbf{h}$ in theory. $\theta = \{W^{\phi}, W^{\psi}, W^{\xi}, F, F^s, F^e\}$ (the bias terms are omitted for clarity) are the model parameters. $\mathbf{h} = \{\mathbf{h}^{\phi}, \mathbf{h}^{\psi}, \mathbf{h}^{\xi}\}$ represent all the hidden variables. $H(\cdot)$ is the entropy functional. In variational learning, we try to find parameters that minimize the Kullback-Leibler divergences between $Q(\mathbf{h})$ and the true posterior $P(\mathbf{h}|\mathbf{V}; \theta)$. A naive mean-field approach can be chosen to obtain a fully factorized distribution for $Q(\mathbf{h})$:

$$Q(\mathbf{h}) = \left[\prod_{r=1}^{R} q(h^{\phi})\right] q(h^{\psi})\, q(h^{\xi}), \tag{18}$$

where $q(h_{rj}^{\phi} = 1) = \mu_{rj}^{\phi}$, $q(h^{\psi} = 1) = \mu^{\psi}$, $q(h^{\xi} = 1) = \mu^{\xi}$. $\boldsymbol{\mu} = \{\mu^{\phi}, \mu^{\psi}, \mu^{\xi}\}$ are the parameters of $Q(\mathbf{h})$. Then the lower bound on the log-probability $\log P(\mathbf{V}; \theta)$ has the form:

$$\begin{aligned} \log P(\mathbf{V}; \theta) \geq &- \bar{E}_{\phi}(\mathbf{V}, \mu^{\phi}) - \bar{\mathbf{E}}_{\psi}(\mathbf{V}, \mu^{\psi}) - \bar{\mathbf{E}}_{\xi}(\mathbf{V}, \mu^{\xi}) \\ &- C(\mathbf{V}) - \bar{E}_{\Phi}(\mu^{\phi}, \mu^{\phi}) - \log \mathcal{Z}, \end{aligned} \tag{19}$$

where $\bar{E}_{\phi}(\mathbf{V}, \mu^{\phi})$, $\bar{E}_{\psi}(\mathbf{V}, \mu^{\psi})$, $\bar{E}_{\xi}(\mathbf{V}, \mu^{\xi})$, and $\bar{E}_{\Phi}(\mu^{\phi}, \mu^{\phi})$ have the same forms, by replacing $\boldsymbol{\mu}$ with $\mathbf{h}$, as Eqs. (7), (8), (9), and (16), respectively.

We can maximize this lower bound with respect to parameters $\boldsymbol{\mu}$ for fixed θ, and obtain the mean-field fixed-point equations:

$$\mu_{rj}^{\phi} = \frac{exp(\sum_{k=1}^{K} W_{jk}^{\phi} \hat{v}_{rk} + D_r a_j^{\phi} + D_{prev}^{j} + D_{next}^{j} - 1)}{\sum_{j'=1}^{J} exp(\sum_{k=1}^{K} W_{j'k}^{\phi} \hat{v}_{rk} + D_r a_{j'}^{\phi} + D_{prev}^{j'} + D_{next}^{j'} - 1)}, \tag{20}$$

$$\mu^{\psi} = \sigma(\sum_{k=1}^{K} W_k^{\psi} \hat{v}_k + Da^{\psi}) \tag{21}$$

$$\mu^{\xi} = \sigma(\sum_{k=1}^{K} W_k^{\xi} \hat{v}_k + Da^{\xi}), \tag{22}$$

where D_{prev}^{j} and D_{next}^{j} are two terms relevant to the derivative of the RHS of Eq. (19) with respect to μ_{rj}^{ϕ}, defined by

$$D_{prev}^{j} = \begin{cases} F_j^s, & r = 1 \\ \sum_{q=1}^{J} \mu_{r-1,q}^{\phi} F_{qj}, & r > 1 \end{cases}$$

$$D_{next}^{j} = \begin{cases} \sum_{q=1}^{J} F_{jq} \mu_{r+1,q}^{\phi}, & r < R. \\ F_j^e, & r = R \end{cases}$$

The updating of $\boldsymbol{\mu}$ can be carried out iteratively until convergence. Then, $(\mathbf{V}, \boldsymbol{\mu})$ can be considered as a special "state" of HSSM, thus the SAP can be applied to update the model's parameters, θ, for fixed $(\mathbf{V}, \boldsymbol{\mu})$.

4 Experiments and Discussions

It's not easy to evaluate the performance of a dialogue structure analysis model. In this study, we examined our model via qualitative visualization and quantitative analysis as done in (Ritter et al., 2010; Zhai and Williams, 2014). We implemented five conventional models to conduct an extensive comparing study on the two corpora: Twitter-Post and AirTicketBooking. Conventional models include: LMHMM (Chotimongkol, 2008), LMHMMS (Ritter et al., 2010), TMHMM, TMHMMS, and TMHMMSS (Zhai and Williams, 2014). In our experiments, for each corpus we randomly select 80% dialogues for training, and use the rest 20% for testing. We select three different number (10, 20 and 30) of latent states to evaluate all the models. In TMHMM, TMHMMS and TMHMMSS, the number of "topics" in the latent states and a dialogue is a hyper-parameter. We conducted a series of experiments with varying numbers of topics, and the results illustrated that 20 is the best choice on the two corpora. So, for all the following experimental results of TMHMM, TMHMMS and TMHMMSS, the corresponding topic configurations are set to 20.

The number of estimation iterations for all the models on training sets is set to 10,000; and on held-out test sets, the numver of iterations for inference is set to 1000. In order to speed-up the learning of HSSM, datasets are divided into mini-batches, each has 15 dialogues. In addition, the learning rate and momentum are set to 0.1 and 0.9, respectively.

4.1 Qualitative Evaluation

Dialogues in Twitter-Post always begin with three latent states: broadcasting what they (Twitter users) are doing now ("Status"), broadcasting an interesting link or quote to their followers ("Reference Broadcast"), or asking a question to their followers ("Question to Followers").[2] We find that structures discoverd by HSSM and LMHMMS with 10 latent states are most reasonable to interpret. For example, after the initiating state ("Status", "Reference Broadcast", or "Question to Followers"), it was often followed a "Reaction" to "Reference Broadcast" (or "Status"), or a "Comment" to "Status", or a "Question" to "Status" ("Reference Broadcast", or "Question to Followers"') etc. Compared with LMHMMS, besides obtaining similar latent states, HSSM exhibits powerful ability in learning sequential dependency relationship between latent states. Take the following simple Twitter dialogue session as an example:

: *rt i like katy perry lt lt we see tht lol*
: *lol gd morning*
: *lol gd morning how u*
: *i'm gr8 n urself*
: *i'm good gettin ready to head out*
: *oh ok well ur day n up its cold out here*
...

LMHMMS labelled the second utterance ("*lol gd morning* ") and the third utterance ("*lol good morning how u* ") into the same latent state, while HSSM treats them as two different latent states (Though they both have almost the same words). The result is reasonable: the first "*gd morning*" is a greeting, while the second "*gd morning*" is a response.

For AirTicketBooking dataset, the state-transition diagram generated with our model under the setting of 10 latent states is presented in Figure 4. And several utterance examples corresponding to the latent staes are also showed in Table 2. In general, conversations begin with sever agent's short greeting, such as "Hi, very glad to be of service.", and then transit to checking the passenger's identity information or

[2] For simplicity and readability in consistent, we follow the same latent state names used in (Ritter et al., 2010)

inquiring the passenger's air ticket demand; or it's directly interrupted by the passenger with booking demand which is always associated with place information. After that, conversations are carried out with other booking related issues, such as checking ticket price or flight time.

The flowchart produced by HSSM can be reasonably interpreted with knowledge of air ticket booking domain, and it most consistent with the agent's real workflow of the Ticket Booking Corporation[3] compared with other models. We notice that conventional models can not clearly distinguish some relevant latent states from each other. For example, these baseline models always confound the latent state "Price Info" with the latent state "Reservation", due to certain words assigned large weights in the two states, such as "打折 (discount)", and "信用卡 (credit card)" etc. Furthermore, Only HSSM and LMHMMS have dialogue specific topics, and experimental results illustrate that HSSM can learn much better than LMHMMS which always mis-recognize corpus general words as belonging to dialogue specific topic (An example is presented in Table 3).

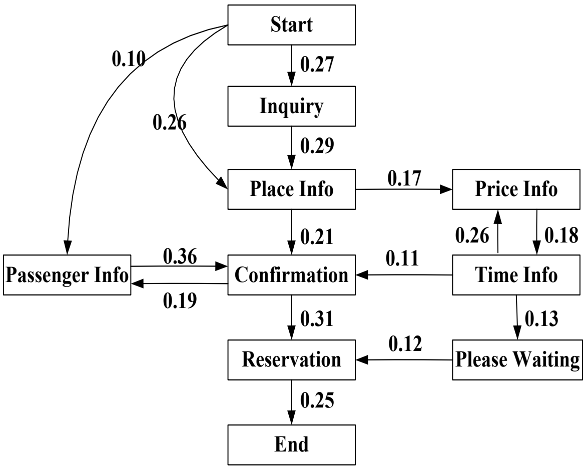

Figure 4: Transitions between latent states on AirTicketBooking generated by our HSSM model under the setting of $J = 10$ latent states. Transition probability cut-off is 0.10.

4.2 Quantitative Evaluation

For quantitative evaluation, we examine HSSM and traditional models with log likelihood and an ordering task on the held-out test set of Twitter-Post and AirTicketBooking.

Latent States	**Utterance Examples (Chinese)**	**Utterance Examples (English Translation)**
Start	您好，很高兴为您服务。	Hello, very glad to be of service.
Inquiry	您想预定机票吗?	Do you want to make a flight reservation?
Place Info	我想预定一张北京到上海的机票。	I want to book an air ticket from Beijing to Shanghai.
Time Info	明天上午10点左右。	Tomorrow morning, about 10 o'clock.
Price Info	成人机票1300元一张。	The adult ticket is 1300 Yuan.
Passenger Info	姓名李东，身份证号12345。	My name is Li Dong, and my ID number is 12345.
Confirmation	好的，可以。	Yes, that's OK.
Please Waiting	请稍等，我帮您查询。	Please wait a moment, I'll check for you.
Reservation	请预定一张，我想用信用卡支付。	Please make a reservation, I want to use a credit card to pay.
End	欢迎下次来电，再见。	Welcome to call next time. Bye.

Table 2: Utterance examples of latent states discovered by our model.

Model	**Top Words**
HSSM	十点, 李东, 福州, 厦门, 上航, ... ten o'clock, Dong Li (name), Fuzhou (city), Xiamen (city), Shanghai Airlines, ...
LMHMMS	有, 十点, 额, 李东, 预留, ... have, ten o'clock, er, Dong Li (name), reserve, ...

Table 3: One example of dialogue specific topic learned on the same dialogue session with HSSM and LMHMMS, respectively.

Log Likelihood The likelihood metric measures the probability of generating the test set using a specified model. The likelihood of LMHMM and TMHMM can be directed computed with the forward algorithm. However, since likelihoods of LMHMMS, TMHMMS and TMHMMSS are intractable to compute due to the local dependencies with respect to certain latent variables, Chib-style estimating algorithms (Wallach et al., 2009) are employed in our experiments. For HSSM, the partition function is a key problem for calculating the likelihood, and it can be effectively estimated by Annealed Importance Sampling (AIS) (Neal, 2001; Salakhutdinov and Murray, 2008).

Figure 5 presents the likelihood of different models on the two held-out datasets. We can observe that HSSM achieves better performance on likelihood than all the other models under different number of latent states. On Twitter-Post dataset our model slightly surpasses LMHMMS, and it performs much better than all traditional models on AirTicketBooking dataset.

Ordering Test Following previous work (Barzilay and Lee, 2004; Ritter et al., 2010; Zhai and Williams, 2014), we utilize Kendall's τ (Kendall, 1938) as evaluation metric, which measures the similarity between any two sequential data and ranges from -1 (indicating a reverse ordering) to $+1$ (indicating an identical

[3] We hide the corporation's real name for privacy reasons.

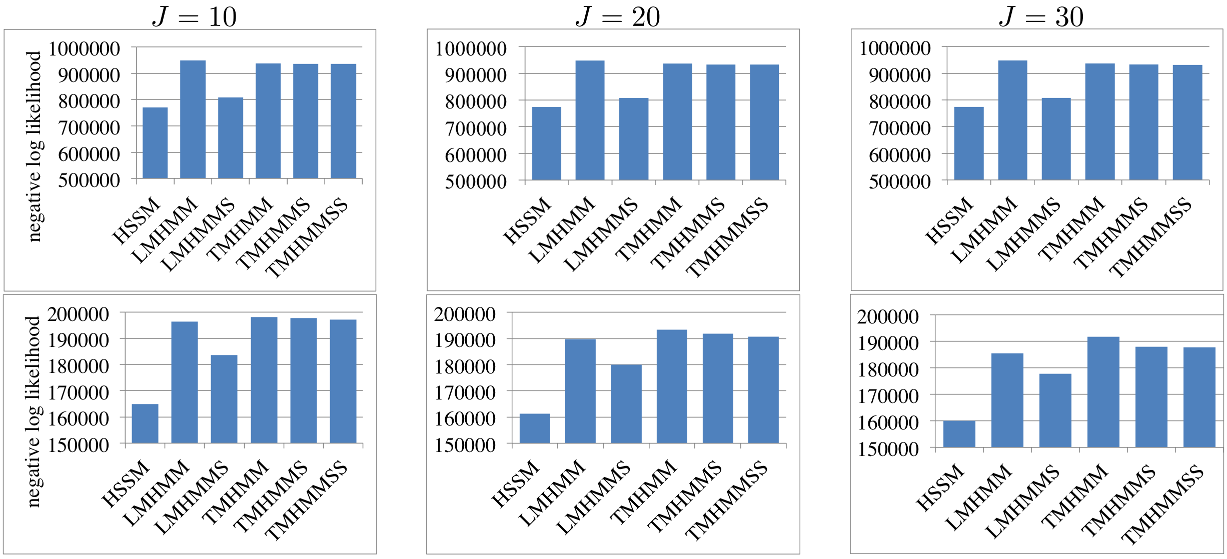

Figure 5: Negative log likelihood (smaller is better) on held-out datasets of Twitter-Post (upper) and AirTicketBooking (lower) under different number of latent states J.

ordering). This is the basic idea: for each dialogue session with n utterances in the test set, we firstly generate all $n!$ permutations of the utterances; then evaluate the probability of each permutation, and measure the similarity, i.e. Kendall's τ, between the max-probability permutation and the original order; finally, we average τ values for all dialogue sessions as the model's ordering test score. As pointed out by Zhai et al. (2014), it's however infeasible to enumerate all possible permutations of dialogue sessions when the number of utterances in large. In experiments, we employ the incrementally adding permutation strategy, as used by Zhai et al. (2014), to build up the permutation set. The results of ordering test are presented in Figure 6. We can see that HSSM exhibits better performance than all the other models. For the conventional models, it is interesting that LMHMMS, TMHMMS and TMHMMSS achieve worse performances than LMHMM and TMHMM. This is likely because the latter two models allow words to be emitted only from latent states (Zhai and Williams, 2014), while the former three models allow words to be generated from additional sources. This also implies HSSM's effectiveness of modeling distinct information uderlying dialogues.

4.3 Discussion

The expermental results illustrate the effectiveness of the proposed undirected dialogue structure analysis model based on Boltzmann machine. The conducted experiments also demonstrate that undirected models have three main merits for text modeling, which are also demonstrated by Hinton and Salakhutdinov (2009), Srivastava et al. (2013) through other tasks. Boltzmann machine based undirected models are able to generalize much better than traditional directed generative model; and model learning is more stable. Besides, an undirected model is more suitable for describing complex dependencies between different kinds of variables.

We also notice that all the models can, to some degree, capture the sequential structure in the dialogues, however, each model has a special characteristic which makes itself fit a certain kind of dataset better. HSSM and LMHMMS are more appropriate for modeling the open domain dataset, such as Twitter-Post used in this paper, and the task-oriented domain dataset with one relatively concentrated topic in the corpus and special information for each dialogue, such as AirTicketBooking. As we known, dialogue specific topics in HSSM or LMHMMS are used and trained only within corresponding dialogues. They are crucial for absorbing certain words that have important meaning but do not belongs to latent states. In addition, for differet dataset, dialogue specific topics may have different effect to the modeling. Take the Twitter-Post for an example, dialogue specific topics formulate actual themes of dialogues, such as a pop song, a sport news. As for the AirTicketBooking dataset, dialogue specific

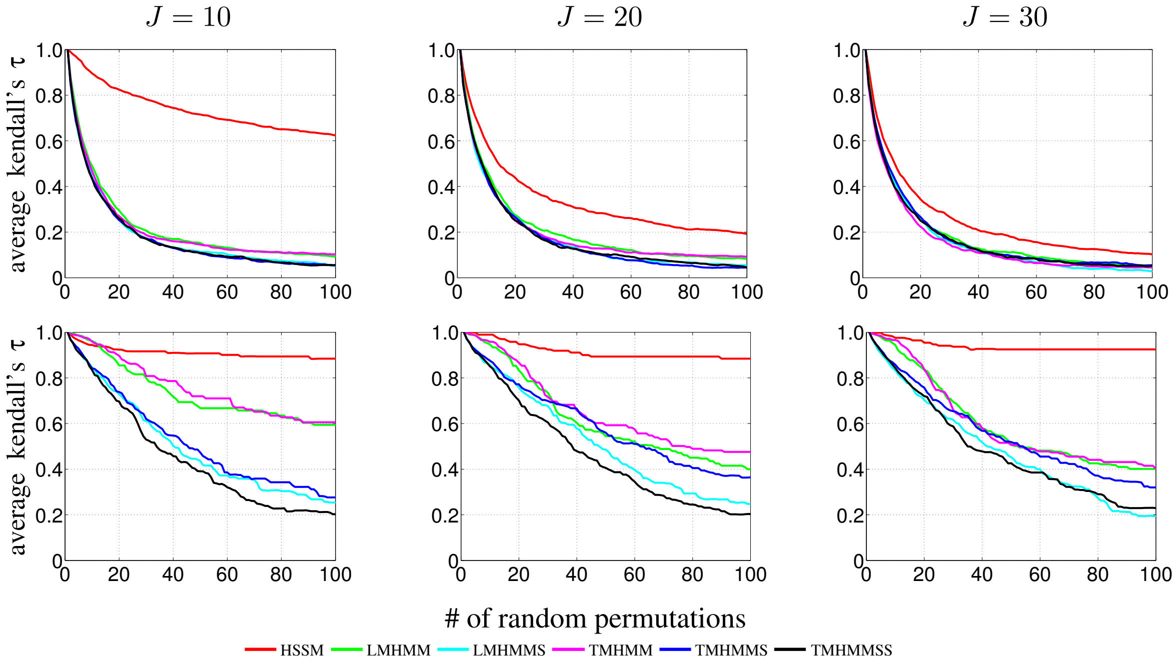

Figure 6: Average Kendall's τ measure (larger is better) on held-out datasets of Twitter-Post (upper) and AirTicketBooking (lower) under different number of latent states J.

topics always represent some special information, such as the personal information, including name, phone number, birthday, etc. In summary, each dialogue specific topic reflects special information which is different from other dialogues.

The three models, TMHMM, TMHMMS and TMHMMSS, which do not include dialogue specific topics, should be utilized on the task-oriented domain dataset, in which each dialogue has little special or personnal information. For example, the three models perform well on the the BusTime and TechSupport datasets (Zhai and Williams, 2014), in which name entities are all replaced by different semantic types (e.g. phone numbers are replaced by "<phone>", E-mail addresses are replaced by "<email>", etc).

5 Conclusions

We develope an undirected generative model, HSSM, for dialogue structure analysis, and examine the effectiveness of our model on two different datasets, Twitter posts occurred in open-domain and task-oriented dialogues from airline ticket booking domain. Qualitative evaluations and quantitative experimental results demonstrate that the proposed model achieves better performance than state-of-the-art approaches. Compared with traditional models, the proposed HSSM has more powerful ability of discovering structures of latent states and modeling different word sources, including latent states, dialogue specific topics and global general topic.

According to recent study (Srivastava et al., 2013), a deep network model exhibits much benefits for latent variable learning. A dialogue may actually have a hierarchy structure of latent states, therefore the proposed model can be extended to a deep model to capture more complex structures. Another possible way to extend the model is to consider modeling long distance dependency between latent states. This may further improve the model's performance.

Acknowledgments

We are grateful to anonymous reviewers for their helpful comments and suggestions. We would like to thank Alan Ritter for kindly providing the raw Twitter dataset.

This work is supported in part by the National Natural Science Funds of China under Grant 61170197 and 61571266, and in part by the Electronic Information Industry Development Fund under project "The R&D and Industrialization on Information Retrieval System Based on Man-Machine Interaction with Natural Speech".

References

James Allen, Nathanael Chambers, George Ferguson, Lucian Galescu, Hyuckchul Jung, Mary Swift, and William Taysom. 2007. Plow: A collaborative task learning agent. In *Proceedings of the National Conference on Artificial Intelligence*, volume 22, page 1514. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models with applications to generation and summarization. *In proceedings of HLT-NAACL 2004*, pages 113–120.

Ananlada Chotimongkol. 2008. *Learning the structure of task-oriented conversations from the corpus of in-domain dialogs*. Ph.D. thesis, SRI International.

William W Cohen, Vitor R Carvalho, and Tom M Mitchell. 2004. Learning to classify email into“speech acts”. In *EMNLP*, pages 309–316.

Nigel Crook, Ramon Granell, and Stephen Pulman. 2009. Unsupervised classification of dialogue acts using a dirichlet process mixture model. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 341–348. Association for Computational Linguistics.

Geoffrey E Hinton and Ruslan R Salakhutdinov. 2009. Replicated softmax: an undirected topic model. In *Advances in neural information processing systems*, pages 1607–1614.

Geoffrey E Hinton and Richard S Zemel. 1994. Autoencoders, minimum description length, and helmholtz free energy. *Advances in neural information processing systems*, pages 3–3.

Geoffrey E Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Minwoo Jeong, Chin-Yew Lin, and Gary Geunbae Lee. 2009. Semi-supervised speech act recognition in emails and forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1250–1259. Association for Computational Linguistics.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. 1999. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Jingjing Liu, Stephanie Seneff, and Victor Zue. 2010. Dialogue-oriented review summary generation for spoken dialogue recommendation systems. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 64–72. Association for Computational Linguistics.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2006. Incorporating speaker and discourse features into speech summarization. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 367–374. Association for Computational Linguistics.

Radford M Neal and Geoffrey E Hinton. 1998. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer.

Radford M Neal. 2001. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations.

Ruslan Salakhutdinov and Iain Murray. 2008. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th international conference on Machine learning*, pages 872–879. ACM.

Nitish Srivastava, Ruslan R Salakhutdinov, and Geoffrey E Hinton. 2013. Modeling documents with deep boltzmann machines. *UAI*.

Tijmen Tieleman. 2008. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071. ACM.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112. ACM.

Yorick Wilks. 2006. Artificial companions as a new kind of interface to the future internet.

Steve J Young. 2006. Using pomdps for dialog management. In *SLT*, pages 8–13.

Ke Zhai and Jason D Williams. 2014. Discovering latent structure in task-oriented dialogues. In *ACL (1)*, pages 36–46.