A Weakly Supervised Method for Topic Segmentation and Labeling in Goal-oriented Dialogues via Reinforcement Learning

Ryuichi Takanobu[1], **Minlie Huang**[1*], **Zhongzhou Zhao**[2],
Fenglin Li[2], **Haiqing Chen**[2], **Xiaoyan Zhu**[1], **Liqiang Nie**[3]
[1] Conversational AI Group, AI Lab., Dept. of Computer Science, Tsinghua University,
Beijing National Research Center for Information Science and Technology, China
[2] Alibaba Group, Hangzhou, China
[3] Shandong University, Jinan, China
gxly15@mails.tsinghua.edu.cn, *aihuang@tsinghua.edu.cn

Abstract

Topic structure analysis plays a pivotal role in dialogue understanding. We propose a reinforcement learning (RL) method for topic segmentation and labeling in goal-oriented dialogues, which aims to detect topic boundaries among dialogue utterances and assign topic labels to the utterances. We address three common issues in the goal-oriented customer service dialogues: *informality, local topic continuity,* and *global topic structure*. We explore the task in a weakly supervised setting and formulate it as a sequential decision problem. The proposed method consists of a state representation network to address the informality issue, and a policy network with rewards to model local topic continuity and global topic structure. To train the two networks and offer a warm-start to the policy, we firstly use some keywords to annotate the data automatically. We then pre-train the networks on noisy data. Henceforth, the method continues to refine the data labels using the current policy to learn better state representations on the refined data for obtaining a better policy. Results demonstrate that this weakly supervised method obtains substantial improvements over state-of-the-art baselines.

1 Introduction

Analyzing topic structures [Arguello and Rosé, 2006; Du *et al.*, 2017] or discourse relations [Afantenos *et al.*, 2015; Qin *et al.*, 2017] of goal-oriented dialogues such as negotiations and customer service conversations, is important for dialogue understanding [Williams *et al.*, 2017], dialogue generation [Li *et al.*, 2016], and dialogue summarization [Bokaei *et al.*, 2016]. In this paper, we focus on analyzing topic structures in goal-oriented dialogues, aiming to detect topic boundaries among utterances and assign topic labels to dialogue utterances[1].

Product-info	**A:**	The release date of ⟨ MODEL ⟩???
	B:	⟨ MODEL ⟩ will be available for pre-order on 19 April and launch on 26.
	A:	How long can the battery last?
	B:	It's equipped with a 4,000 mAh battery up to 8 hours of HD video playing or 10 hours of web browsing.
Promotion	**A:**	Can I use a coupon?
	B:	When entering your payment on the checkout page, click *Redeem a coupon* below your payment method.
	B:	You can check here for more details: ⟨ URL ⟩.
Payment	**A:**	OK. Support payment by installments?
	B:	Sure. We provide an interest-free installment option for up to 6 months.

Table 1: An example of customer service dialogues, translated from Chinese. Utterances in the same color are of the same topic.

Different from other generic text, goal-oriented dialogues have the following three distinctive features. 1) **informality** : a user may post fragmented, incomplete sentences with typos, colloquialisms, or informal terms, particularly in customer service dialogues (see Table 1). 2) **local topic continuity**: it usually takes several turns to discuss one topic and it maintains the same topic until the current problem has been resolved. 3) **global topic structure**: a dialogue session has clear boundaries, few cross-transitions between topic segments, but high cohesion within one segment.

However, existing methods cannot fully address the aforementioned issues. Many methods capture local topic continuity by employing local lexical cohesion based on word or phrase similarity [Purver *et al.*, 2006; Eisenstein and Barzilay, 2008]. They do not consider the sentence-level dependencies and are unable to appropriately summarize the context. Moreover, they are weak to make coherent local topic assignment, and generally produce fragmented seg-

*Corresponding author: Minlie Huang.

[1]Topic in dialogues can be viewed as coarse-grained intent and thus topic analysis offers intent understanding to some degree.

ments. Other studies attempted to capture the discourse dependencies between adjacent utterances [Du *et al.*, 2017; Zhai and Williams, 2014], but less attention has been paid to modeling global topic structure in a dialogue session. Fully supervised methods [Arguello and Rosé, 2006], by contrast, are too expensive for manual annotation, thereby not scalable to large datasets.

In this paper, we propose a policy gradient reinforcement learning method to address the three issues. Topic segmentation and labeling can be seen as a sequential decision problem, where we assign topics sequentially to utterances, and previous decisions can affect current and following decisions. An intermediate reward is defined to encourage local topic continuity to enforce the coherence of local topic assignment from the *labeling perspective*. When all sequential decisions are made, the global topic structure of one session is measured by a delayed reward that favors larger utterance similarity within a segment and lower similarity between segments. To address the informality issue, we use a hierarchical LSTM (HLSTM) for state representation to capture word-level and sentence-level dependencies. HLSTM can better summarize all historical information instead of just using word/phrase similarity. It thus has the ability of context understanding to not only deal with informality but also address local topic continuity from the *content perspective*.

Our method consists of a state presentation network and a policy network. As can be imagined, state representations are extremely important to our method, however, without labeled data, it is challenging to learn a good representation of text. Another challenge for the policy network is its inability of topic labeling with the designed rewards alone. In order to train state representations and also provide a warm-start for the policy to identify topics, labeled data are indispensable. Unfortunately, it is too costly to manually annotate the large-scale data in our task. Although some unsupervised methods can assign latent topics [Blei, 2012], however, such topics are indirect and lack direct interpretability for the task. Therefore, we resort to *noisy labeling*, using a set of handcrafted keywords to label the dialogues automatically. After pre-training with the noisy data, the method continues to refine the data labels using the current policy, and then to learn better state representations on the refined data for obtaining a better policy. To summarize, our contributions are as follows:

- We propose a weakly supervised method for analyzing topic structures of goal-oriented dialogues. To avoid heavy manual annotation, we use prior knowledge to perform noisy labeling for pre-training the networks. The method iterates between refining noisy data labels and finding better state representations and policies. Thus, it is scalable to large unlabeled datasets and the idea may inspire other real-world applications.
- Our method is able to capture *local topic continuity* by an intermediate reward, measure *global topic structure* by a delayed reward, and represent *dialogue content and context* by a hierarchical LSTM. It generates not only locally coherent but also globally well-structured topic segments. Experimental results demonstrate substantial improvements over the baselines.

2 Related Work

Early models for topic segmentation assumed that a high lexical cohesion is expected within a topic segment [Hearst, 1997]. However, most of those models put much emphasis on the lexical structures of a dialogue [Webber *et al.*, 2012]. Such superficial signals like words or phrases do not consider sentence-level dependencies, leading to a fragmented segmentation. Other studies aimed to discover latent discourse structure by modeling message-response pairs [Du *et al.*, 2017] or drawing a distribution over topic-state links [Zhai and Williams, 2014], but the global structure of a session is often ignored. Though TopicTilling [Riedl and Biemann, 2012] used latent topics obtained from LDA to represent sentences, such topics are still far from an appropriate summary of the context. While few supervised approaches cast the segmentation problem as a classification task [Arguello and Rosé, 2006], it is obviously expensive for annotation and not scalable to large datasets. Recent research [Song *et al.*, 2016] has demonstrated that word embedding provides better performance in dialogue session segmentation, which inspires us to apply a state representation network for better summarizing and understanding dialogue contexts.

Topic labeling aims to assign a short description to each of the topical clusters to facilitate interpretations of the topics [Joty *et al.*, 2013]. In general, single terms [Mehrotra *et al.*, 2013] or phrases [Mei *et al.*, 2007] can be chosen as topic label. Most researchers formulated the task as a multi-label classification problem that each utterance is associated with a subset of all the topics, and multiple topic labels may be chosen for an utterance [Ramage *et al.*, 2009; Soleimani and Miller, 2016]. In this paper, we treat topic labeling as a topic classification problem where each utterance is assigned with a topic in a given set.

There are a few challenges for topic segmentation and labeling in goal-oriented dialogues. First, participants are discussing about very specific issues whose topics are restricted in certain domains, but existing unsupervised methods have very limited abilities to learn domain-specific knowledge from the dataset [Joty *et al.*, 2013]. Second, dialogue messages are ungrammatical and informal, where parsing tools are not applicable [Du *et al.*, 2017]. To address these two problems, we cast topic segmentation and labeling in dialogues as a RL problem and design rewards to model local topic continuity and global topic structure.

3 Methodology

Our task is to segment a dialogue session and label each segment with a topic. More formally, given a sequence of dialogue utterances $X = x_1, x_2, \cdots, x_T$ where each x_i is an utterance, and a topic set $\mathcal{C}$ as $\{c_1, c_2, \cdots, c_K\}$, the task is to assign each x_i with a topic c_j. We treat the task as a topic classification problem in which topics have been specified in advance. [2] Note that we do not consider speaker turns, since utterances from a customer or an agent are in the same topic space.

[2]The task definition differs from other studies [Joty *et al.*, 2013], whereby topic labeling is defined as finding a short description to each of the topical clusters to facilitate topic interpretations.

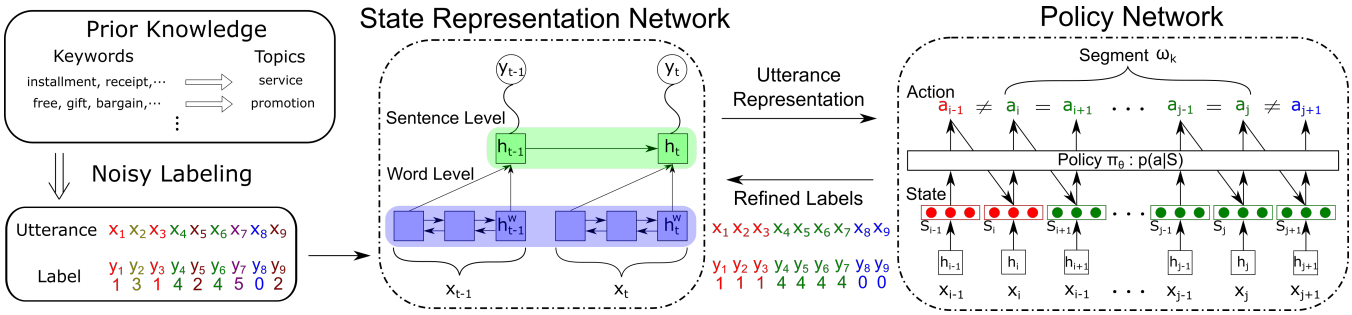


Figure 1: Illustration of the model. SRN adopts a hierarchical LSTM to represent utterances and provides state representations to PN. Data labels are refined to retrain SRN and PN to learn better state representations and policies. The label y and the action a are in the same space.

Our model, as illustrated in Figure 1, consists of a state representation network (SRN) and a Policy Network (PN). We use a set of keywords (as prior knowledge) to perform noisy labeling on the dialogues, and pre-train SRN and PN on the noisy data. SRN is a hierarchical LSTM (HLSTM) consisting of a word-level LSTM and a sentence-level LSTM, which fully captures word-level and sentence-level context dependencies. Since noisy labeling may not assign correct labels to the context-dependent utterances, PN will refine topic labeling by optimizing the cumulative rewards that model local topic continuity and global topic structure. The refined data are used to train SRN again to provide better state representations and then obtain better policies for PN. The process iterates until it converges.

3.1 Noisy Labeling with Prior Knowledge

Prior knowledge in this paper is defined as a set of keywords for each topic and is used to label the dialogues automatically. The noisy data is used to pre-train state representations and also to provide a warm-start for the policy network. These keywords are manually selected from a ranking list of words. Compared to manual annotation, such prior knowledge can be obtained more easily.

The annotation process works as follows: **First**, we use the strategy of *keyword matching (KM)*. If an utterance contains the keywords of a topic, the utterance will be assigned with the topic accordingly. Formally, for an utterance x, its topic label $l(x)$ is decided by

$$l(x) = \text{argmax}_j \sum_i tf(k_{ij}, x), \tag{1}$$

where k_{ij} denotes the i-th keyword for topic j, and $tf(k, x)$ counts the frequency of keyword k in x.

Second, we adopt the *nearest neighbouring (1-NN)* strategy if there is no keyword occurring in an utterance. We compute the inner product of the topic vector and the utterance vector, and choose the most similar topic. The topic vector and utterance vector are both the average of word vectors.

3.2 State Representation Network (SRN)

We adopt a hierarchical LSTM (HLSTM)[Chung *et al.*, 2017] to offer state representations to the policy network (PN). A word level bidirectional LSTM connects words in an utterance forward and backward, and the utterance representation is fed into the sentence-level LSTM which sequentially connects utterances in a dialogue session. More formally, given a sequence of utterances $X = x_1, x_2, \cdots, x_L$, the sentence-level LSTM obtains the hidden states $\mathcal{H} = \{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_L\}$ from which state representations will be computed for PN. Such a HLSTM can capture word-level and sentence-level dependencies. For instance, an informal word can be better summarized by its contexts.

SRN is pre-trained using the noisy data. We adopt cross entropy to train SRN where supervision is imposed on the topic prediction (y_{t-1} and y_t in Figure 1) based on the hidden states of the sentence-level LSTM, i.e., $\mathbf{h}_{t-1}$ and $\mathbf{h}_t$. SRN will be fine-tuned during the joint training of SRN and PN.

3.3 Policy Network (PN)

The policy network adopts a stochastic policy which is a probability distribution over actions given a state, $\pi_\Theta(a_t|\mathbf{S}_t)$. State representation is composed of the output of SRN and the latest topic segment. The action corresponds to assigning a topic label to an utterance in a dialogue session. We design an intermediate reward to capture local topic continuity since the topics in a session usually last for several turns, and a long-term delayed reward to capture the global clustering property of topic segments: the intra-segment similarity is large while the inter-segment similarity is small.

Action

The action space is the same as the topic space, i.e., $\mathcal{A} = \{c_1, c_2, \cdots, c_K\}$. Thus, at each state, the action is to assign a topic label to the current utterance. A topic boundary can be immediately identified when the topics of two adjacent utterances are different.

State

The state $\mathbf{S}_t$ represents the current state after the actions $a_1, a_2, ..., a_{t-1}$ are sampled for the previous utterances in a dialogue session. The state can be formally represented by

$$\mathbf{S}_t = [\mathbf{h}_t; \mathbf{v}(a_{t-1}); \boldsymbol{\omega_k}], \tag{2}$$

where $\mathbf{v}(a)$ denotes the vector of topic label a, a learnable parameter of PN, and $\boldsymbol{\omega_k}$ is the vector representation of the latest segment ω_k, defined as $\boldsymbol{\omega} = \frac{1}{|\omega|} \sum_{x \in \omega} \mathbf{h}_x$.

Policy

A stochastic policy π_Θ is adopted to sample a topic label for the current utterance at the current state $\mathbf{S}_t$. The policy function is defined by a softmax function, which is a probability distribution over all topic labels in $\mathcal{C}$ ($\mathcal{A} = \mathcal{C}$). As the probability of each topic differs from one to another, the policy tends to choose the topic with a larger probability even though other choices may derive the same reward. We thus extend the softmax function with a $\beta - smoothed$ version, which has been reported to be effective in recent research [Guu *et al.*, 2017].

$$\pi_\Theta(a_t|\mathbf{S}_t) = softmax(\mathbf{z};\beta) = \frac{\exp(\mathbf{z}\cdot\beta)}{\sum_{\tilde{\mathbf{z}}}\exp(\tilde{\mathbf{z}}\cdot\beta)}, \quad (3)$$

where $\mathbf{z} = \mathbf{W}\cdot\mathbf{S}_t$ is the input to the softmax function. Obviously, a very large β leads to the effect of $argmax$ while smaller values produce more smoothed distributions.

Reward

We design two rewards: one is an intermediate reward to capture the local continuity of topics in dialogues since one topic usually lasts for several turns and it will not change too frequently; the other is a long-term delayed reward that models the global clustering property of topic segments: the content similarity between segments should be small while the similarity within a segment should be large.

The intermediate reward encourages the continuity of local topics during interactions. Formally

$$r_{int} = \frac{1}{L-1} sign(a_{t-1} = a_t)\cos(\mathbf{h}_{t-1}, \mathbf{h}_t),$$

where $sign(c)$ is 1 if the condition c is true and -1 otherwise, and $cos(\cdot)$ is the cosine similarity between two vectors. This reward addresses not only local topic continuity from the *labeling perspective*, but also the content similarity between adjacent utterances from the *content perspective*, where the content $(\mathbf{h}_{t-1}, \mathbf{h}_t)$ is represented by SRN.

The long-term delayed reward can be obtained when all utterances in a session are assigned with topic labels. To obtain a good global topic structure, the reward encourages higher similarity between utterances within the same segment and lower similarity between adjacent segments, as below:

$$r_{delayed} = \frac{1}{N}\sum_{\omega\in X}\frac{1}{|\omega|}\sum_{X_t\in\omega}\cos(\mathbf{h}_t, \boldsymbol{\omega}) - \frac{1}{N-1}\sum_{(\omega_{k-1},\omega_k)\in X}\cos(\boldsymbol{\omega}_{k-1}, \boldsymbol{\omega}_k),$$

where N is the number of segments predicted by the policy in a dialogue session X, ω is a topic segment in X, and $\boldsymbol{\omega}$ is its vector representation. Note that topic segments can be obtained only when the topics of all utterances are sampled by PN, therefore, we call the reward *delayed*.

Training

We use policy gradient methods [Sutton *et al.*, 2000] with the REINFORCE algorithm [Williams, 1992] for optimization, aiming to maximize the expected total reward for a dialogue session. Denote τ as a sequence $\mathbf{S}_1, a_1, \cdots \mathbf{S}_L, a_L$ generated from the policy, and $\mathcal{T}$ as the set of all possible sequences. The expected reward for a dialogue session $X = x_1, x_2, \cdots, x_L$ can be computed as follows:

$$J(X;\Theta) = \mathbb{E}_{(\mathbf{S}_t,a_t)\sim P_\Theta(\mathbf{S}_t,a_t)}[\sum_{t=1}^{L} r(\mathbf{S}_t, a_t)] = \sum_{\tau\in\mathcal{T}} p(\mathbf{S}_1)\prod_{t=1}^{L}\pi_\Theta(a_t|\mathbf{S}_t)p(\mathbf{S}_{t+1}|\mathbf{S}_t, a_t)R_L,$$

where $R_L = (\sum_{t=1}^{L} r_{int}) + r_{delayed}$ is the expected cumulative reward for a dialogue session. Then, the gradient is estimated using the likelihood ratio trick:

$$\nabla_\Theta J(X;\Theta) = \mathbb{E}_{\tau\sim\pi_\Theta(\tau)}[\sum_{t=1}^{L}(R_L - b(\tau))\nabla_\Theta \log\pi_\Theta(a_t|\mathbf{S}_t)]. \quad (4)$$

The *baseline* $b(\tau)$ [Williams, 1992] in Eq.4 is used to reduce the variance of the estimate without altering its expectation theoretically. In practical use, we will sample some sequences $\tau_1, \tau_2 \cdots \tau_k$ for X with the current RL policy. The model will assign a score to each sequence according to the designed reward function, and then estimates $b(\tau)$ as the average of those rewards. So the coefficient $R_L - b(\tau)$ will be positive if the reward of sampled sequence τ is larger than $b(\tau)$, to encourage a good exploration sampled by RL policy, otherwise negative.

The training details of PN is summarized in Algorithm 1.

Algorithm 1: Reinforcement Learning Process

```
Require: D as training data;
foreach dialogue session X_{1:L} ∈ D do
    Initialize a_0 and ω_k with zeros;
    for t ← 1 to L do
        Obtain state S_t with h_t, a_{t-1}, and ω_k;
        Sample action a_t ~ π_Θ(a_t|S_t) by Eq.(3);
        Update ω_k with a_t and a_{t-1};
    end
    Compute delayed reward R_L;
    Θ ← Θ + α∇J(Θ) using Eq.(4);
end
```

3.4 Overall Procedure

The overall procedure is shown in Algorithm 2. First of all, the training and validation dialogues are labeled by prior

Algorithm 2: Overall Training Process

1. Initialize training data D, validation data V by noisy labeling;
2. Pre-train SRN and PN on D ;
3. Obtain refined data D', V' for all sessions in D, V by applying the current policy: $a_t^* = argmax_a \pi_\Theta(a|\mathbf{S}_t)$;
4. Train SRN on D' to update state representations of D', V';
5. $D \leftarrow D'$, $V \leftarrow V'$;
6. Train PN on D with Algorithm 1;
7. Repeat Step 3-6 until the relative change ratio (RCR, see the section of *Convergence Analysis*) between V and V' at Step 3 is less than 0.5%;

knowledge (a set of keywords). SRN and PR are then pre-trained on the noisy data to provide a warm-start. The algorithm then begins the iterative procedure: apply the current policy to refine data labels by considering local topic continuity and global topic structure, train SRN on the refined data to obtain better state representations, and then train PN to obtain better policies. Since better policies can further refine noisy data labels, the algorithm runs in a positive loop until it converges.

4 Experiment

4.1 Datasets

Due to the lack of benchmark dialogue datasets of topic segmentation and labeling, we collected customer service dialogues from a large E-commerce website. The dialogues came from two domains: SmartPhone and Clothing. These dialogues recorded the interactions between the customers and the merchant agents on products inquiry, delivery services, and other related information.

Datasets	SmartPhone	Clothing
# Topic category	7	10
# Training session	12,315	10,000
# Training utterance	430,462	338,534
# Gold-standard session	300	315
# Gold-standard utterance	10,888	10,962

Table 2: Statistics of the corpus.

Category	Keywords list
Service	分期(installment),发票(receipt), 额度(quota),电子(electronic)
Promotion	送(free),活动(promotion),礼(gift), 赠品(free gift),享(share)
Chitchatting	嗯(yeah),谢谢(thank),你好(hello), 恩恩(yep),客气(welcome)
Product Info	卡(card),膜(film),耳机(headset), 款(model),充电(charge)
Refunds	退款(refund),退(return),运费(fee), 同意(agree),寄回(send back)
Logistics	发货(delivery),快递(express),发(send), 默认(default),预计(estimated)
Payment	改(change),更改(adjust),错(wrong), 留下(record),无效(invalid)

Table 3: Topics and sample keywords in the SmartPhone corpus.

Table 2 details the corpus. An utterance refers to the text from a customer or an agent, which may contain incomplete or multiple sentences. The training data are unlabeled while the test data are annotated manually. We randomly chose 300 sessions from the training dataset for validation. The topics and some sample keywords for each topic in the SmartPhone domain are shown in Table 3. These keywords of all topic categories are manually selected based on frequency and with some heuristics. Similar processes are conducted on the Clothing dataset.

4.2 Experiment Settings

The parameters of SRN are set as follows: the dimension of the hidden states and word vectors is both 100, and the learning rate is 1e-4. In PN, we performed a grid search for hyper-parameters to maximize the total reward $R_V = \sum_{X \in V} J(X; \Theta)$ on the validation set V. These hyper-parameters include: the dimension of the topic vectors = 100, the learning rate α = 1e-5 and β = 0.5 in the smoothed softmax.

4.3 Evaluation Metrics

Our method is evaluated on two tasks: topic segmentation and topic labeling. We adopted the below metrics: (1) Mean absolute error (MAE) and WindowDiff (WD) for topic segmentation; (2) and classification accuracy for topic labeling.

As for MAE, we compared the number of predicted segments in each dialogue session with the gold standard segmentation on the test set T, formally defined as $MAE = \frac{1}{|T|} \sum_{X \in T} |N_{pred}(X) - N_{ref}(X)|$ where $N_{pred}(X)$ denotes the number of segments in X predicted by a model, and $N_{ref}(X)$ the number of reference segments.

WindowDiff (WD) for topic segmentation is adopted from [Pevzner and Hearst, 2002]. WD moves a sliding window of fixed size w over a dialogue session to compare the result predicted by the model with the reference result. $0 \leq WD \leq 1$ with a perfect segmenter scoring 0. As a suggested setting, the window size w is set to 3/4, almost half of the average segment length of 5.02/7.67 utterances per session in the SmartPhone/Clothing domain, respectively.

(a) Topic Segmentation (MAE and WD)

Model	SmartPhone		Clothing	
	MAE	WD	MAE	WD
TextTiling(TeT)	13.09	.802	16.32	.948
TopicTiling	3.30	.522	3.67	.602
TeT+Embedding	3.59	.564	3.17	.567
STM	4.37	.505	8.85	.669
NL+HLSTM	8.25	.632	16.26	.925
Our method	**2.69**	**.415**	**2.74**	**.446**

(b) Topic Labeling (Accuracy)

Model	SmartPhone	Clothing
Twitter-LDA	25.4	24.5
TopicTiling	27.6	17.2
Keyword Matching	39.8	31.8
NL	51.4	39.0
NL+HLSTM	52.6	40.1
Our method	**62.2**	**48.0**

Table 4: Results on topic segmentation (a) and topic labeling (b). NL denotes noisy labeling with prior knowledge.

4.4 Main Results

Topic Segmentation

We compared our method with the following baselines:

▷ *TextTiling* [Hearst, 1997]: It measures the similarity of each

(a)

Model	# Keywords per topic		
	3	6	9
NL	45.0	51.4	48.0
NL+HLSTM	46.6	52.6	48.8
Our method	**55.3**	**62.2**	**58.2**

(b)

SubSets	KM	1-NN
Utterances	3,503	7,385
NL	78.7	38.4
NL+HLSTM	78.6	40.2
Our method	**79.0**	**54.2**

(c)

Model Setting	Segmentation		Labeling
	MAE	WD	Acc
RL + r_{int}	3.04	.449	59.5
RL + $r_{delayed}$	3.89	.490	60.4
RL + r_{int} + $r_{delayed}$	**2.69**	**.415**	**62.2**

Table 5: (a) Labeling accuracy with different numbers of keywords. (b) Labeling accuracy on the subsets of the test set. (c) Influence of different rewards on the performance of segmentation and labeling. All the experiments were conducted on the SmartPhone domain.

adjacent sentence pair, and "valleys" of similarities are detected for segmentation.

▷ *TopicTiling* [Riedl and Biemann, 2012]: It employs latent topics obtained from LDA to improve performance by stabilizing the topics.

▷ *TextTiling+Embedding* [Song *et al.*, 2016]: Embedding enhanced TextTiling, by applying word embeddings to compute similarity between sentences.

▷ *STM* [Du *et al.*, 2013]: A Structured Topic Model, assuming that any utterance in the same segment is generated from the same segment-level topic distribution.

▷ *NL+HLSTM*: The topic label of each utterance is obtained by this supervised model and then topic segments are obtained by merging adjacent utterances of the same topic. The model is trained on the noisy data.

Results are presented in Table 4(a). Lower MAE and WD scores indicate a better agreement with the gold standards. Our method outperforms other models in terms of MAE and WD on both domains. TextTiling and NL+HLSTM tend to make smaller segments compared to other baselines. TopicTiling has good performance on topic segmentation since it tries to find fine-grained subtopical changes using LDA. Embedding enhanced TextTiling works much better since word embedding can well capture the semantic similarity. STM uses a hierarchical Bayesian model by representing utterances with topic distributions. However, no baseline considers global topic structure as our method does.

Topic Labeling

We compared the following baselines for this task:

▷ *Twitter-LDA* [Zhao *et al.*, 2011]: We regard a dialogue session as a text document in LDA, and each utterance is assigned with a topic label using Twitter-LDA. The mapping between a latent topic and a topic label of the task is manually built.

▷ *TopicTiling* [Riedl and Biemann, 2012]: We use topic ID assigned by TopicTiling to label the sentences, and also set up a mapping from a topic ID to a topic label.

▷ *Keyword Matching*: Topic label is assigned by keyword matching, see Eq. 1.

▷ *NL*: Topic label of an utterance is given by noisy labeling with the prior knowledge using both keyword matching and nearest neighboring, see *Section 3.1*.

▷ *NL+HLSTM*: The same as that in topic segmentation. Prediction for each utterance depends on the preceding utterances of the same session.

Results in Table 4(b) show that: 1) Our method outperforms other models substantially indicating that optimizing local and global topic structures can benefit the task greatly; 2) Because the generated latent topics are implicit and indirect, unsupervised methods (Twitter-LDA, TopicTiling) obtain an accuracy lower than 30% on the two datasets. On the contrary, *Keyword Matching* is more straightforward and simple, and can provide a better result; 3) NL+HLSTM performs slightly better than NL as expected. Training a supervised model on noisy data does not lead to better results compared to the noisy labeling since noisy label is the upper bound, indicating that it has a limited ability for context understanding if without the ability to correct noisy labels.

4.5 Extended Evaluation

In the following sections, we evaluated the robustness of our method, the influence of different rewards, and the convergence analysis on the SmartPhone domain.

Robustness to Prior Knowledge

The noisy data labeled by prior knowledge provides a warm-start to state representation and policy learning. To verify whether our method can obtain robust improvements over the baselines, we varied the number of keywords in each topic for noisy labeling. The keywords are ranked by their frequencies in the corpus in descending order.

As shown in Table 5(a), our method can improve the results robustly and substantially even if the number of keywords in prior knowledge is varied. Few keywords (3) may not provide sufficiently good warm-start for our method, but it still has remarkable improvement against the baselines (55.3% vs. 46.6%). Too many keywords (9) degrade the performance remarkably since the discriminative ability between topics decreases, however, our method still has much better results than the baselines (58.2% vs. 48.8%). In all the cases, our method can improve the baselines by absolute values of 10$\sim$12%.

Ability of Context Understanding

To verify the ability of context understanding of our method, we split all the test utterances to two sets: the first subset whose labels can be assigned by keyword matching (*KM*), and the second subset which contains no keyword and is labeled by the nearest neighbor strategy (*1-NN*, see *Section 3.1*). In general, labeling the second subset correctly requires understanding the context, which is harder for the models. The two sets have no intersection.

Results in Table 5(b) show that: **1)** Most utterances (about 68%) do not contain topical keywords. Given that informal

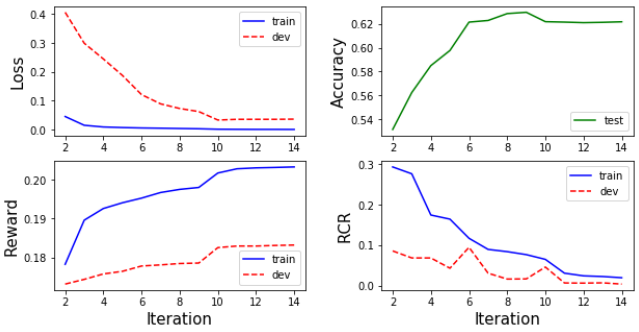


Figure 2: Learning curves of loss/reward (left) and testing accuracy/RCR (right) of PN during iterations. First two iterations are omitted for pre-training.

language is commonly used through dialogues, the second subset is much more difficult than the first one for this task. **2)** Since our method has the ability to correct noisy labels and to represent dialogue content and context by considering all historical information, it outperforms NL and NL+HLSTM substantially on the second subset, and it can even improve the accuracy on the first set. Thus, our method can better understand the context.

Influence of the Rewards

We justified the influence of the intermediate reward and delayed reward on the performance.

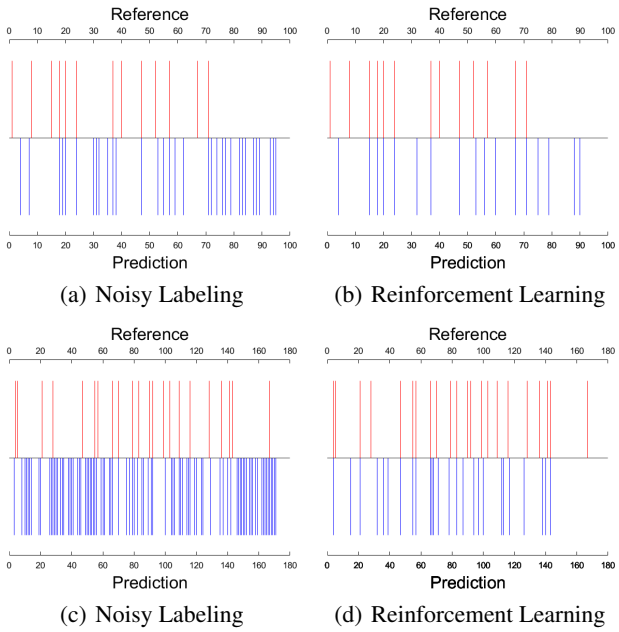
Table 5(c) demonstrates that our method achieves the best performance when optimizing the intermediate and delayed rewards simultaneously. When removing the delayed reward, labeling accuracy drops substantially, indicating that global topic structure is a key factor in this task. When removing the intermediate reward, labeling accuracy also drops remarkably, indicating that local topic continuity also matters. Similarly, the model leads to suboptimal performance on topic segmentation when ablating either reward. As it can be observed, the intermediate reward is more influential on segmentation than the delayed reward, because local topic continuity is a more straightforward factor which manipulates segmentation from the labeling perspective.

Convergence Analysis

The learning curves (loss and reward) in Figure 2 show that our model converges after 14 iterations. The testing accuracy reaches a stable value (62.2%) after the model converges.

In order to verify how the noisy data are changed by the RL module, we proposed relative change ratio (RCR) to quantify the changes of labels between the dataset A (before RL) and A' (after refined by RL) (see Algorithm 2). Define the set $\Delta(A; A') = \{x | y \neq y', (x, y) \in A \wedge (x, y') \in A'\}$ where y is the label of utterance x, and RCR is then calculated as $RCR(A) = \frac{|\Delta(A;A')|}{|A|}$. RCR measures the percentage of utterances with labels changed before and after the RL process.

Figure 2 indicates that the $RCR(V)$ of both the training and validation data converges to a very small percentage (<0.5%) after 14 iterations. Meanwhile the accuracy on the test set is gradually improved. This clearly shows that RL



(a) Noisy Labeling (b) Reinforcement Learning

(c) Noisy Labeling (d) Reinforcement Learning

Figure 3: Exemplar segmentation from the reference segmentation, Noisy labeling (left), and RL (right). The horizonal axis is the index of utterances in a session.

continuously improves the result thanks to the rewards modeling the local and global properties of topic structures.

Topic Structure Visualization

To provide more insights, we visualized the segment results for a session example from noisy labeling, RL, and the gold annotation, respectively. Each bar indicates a topic boundary between utterances in a session. Red bar indicates the result from human annotation, and blue bar from noisy labeling or our RL method.

As can be seen in Figure 3, noisy labeling tends to produce more fragmented segments, while RL can provide much more coherent segmentation.

5 Conclusion

We present a weakly supervised method for topic segmentation and labeling in goal-oriented dialogues. Our central logic works as follows: noisy labeling provides a warm-start to state representation training and policy learning, data refinement can be obtained by optimizing the rewards which capture both local topic continuity and global topic structure, and better data can be used to train better state representations and policies. This positive loop can be run iteratively until the model converges. This methodology can be generalized to other tasks. Through correcting noisy labels of automatically annotated data, a weakly supervised method can improve the performance substantially, if some task/domain/prior expertise can be well captured by the reward function. Extensive experiments show that the method has a strong ability of segmentation, labeling, and context understanding.

Such a method, firstly noisy labeling and then refining with RL, may inspire other tasks to obtain superior performance in

weakly supervised settings.

Acknowledgements

This work was partly supported by the National Science Foundation of China under grant No.61272227/61332007 and the National Basic Research Program (973 Program) under grant No. 2013CB329403.

References

[Afantenos *et al.*, 2015] Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. Discourse parsing for multi-party chat dialogues. In *EMNLP*, pages 928–937, 2015.

[Arguello and Rosé, 2006] Jaime Arguello and Carolyn Rosé. Topic segmentation of dialogue. In *HLT-NAACL Workshop on Analyzing Conversations in Text and Speech*, pages 42–49, 2006.

[Blei, 2012] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[Bokaei *et al.*, 2016] Mohammad Hadi Bokaei, Hossein Sameti, and Yang Liu. Extractive summarization of multi-party meetings through discourse segmentation. *Natural Language Engineering*, 22(1):41–72, 2016.

[Chung *et al.*, 2017] Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. Hierarchical multiscale recurrent neural networks. In *ICLR*, 2017.

[Du *et al.*, 2013] Lan Du, Wray L Buntine, and Mark Johnson. Topic segmentation with a structured topic model. In *HLT-NAACL*, pages 190–200, 2013.

[Du *et al.*, 2017] Wenchao Du, Pascal Poupart, and Wei Xu. Discovering conversational dependencies between messages in dialogs. In *AAAI*, pages 4917–4918, 2017.

[Eisenstein and Barzilay, 2008] Jacob Eisenstein and Regina Barzilay. Bayesian unsupervised topic segmentation. In *EMNLP*, pages 334–343, 2008.

[Guu *et al.*, 2017] Kelvin Guu, Panupong Pasupat, Evan Zheran Liu, and Percy Liang. From language to programs: Bridging reinforcement learning and maximum marginal likelihood. In *ACL*, pages 1051–1062, 2017.

[Hearst, 1997] Marti A Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64, 1997.

[Joty *et al.*, 2013] Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. Topic segmentation and labeling in asynchronous conversations. *Journal of Artificial Intelligence Research*, 47:521–573, 2013.

[Li *et al.*, 2016] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. In *EMNLP*, pages 1192–1202, 2016.

[Mehrotra *et al.*, 2013] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *SIGIR*, pages 889–892, 2013.

[Mei *et al.*, 2007] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *SIGKDD*, pages 490–499, 2007.

[Pevzner and Hearst, 2002] Lev Pevzner and Marti A Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002.

[Purver *et al.*, 2006] Matthew Purver, Thomas L Griffiths, Konrad P Körding, and Joshua B Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. In *ACL*, pages 17–24, 2006.

[Qin *et al.*, 2017] Kechen Qin, Lu Wang, and Joseph Kim. Joint modeling of content and discourse relations in dialogues. In *ACL*, pages 974–984, 2017.

[Ramage *et al.*, 2009] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, pages 248–256, 2009.

[Riedl and Biemann, 2012] Martin Riedl and Chris Biemann. Topictiling: a text segmentation algorithm based on lda. In *ACL Student Research Workshop*, pages 37–42, 2012.

[Soleimani and Miller, 2016] Hossein Soleimani and David J Miller. Semi-supervised multi-label topic models for document classification and sentence labeling. In *CIKM*, pages 105–114, 2016.

[Song *et al.*, 2016] Yiping Song, Lili Mou, Rui Yan, Li Yi, Zinan Zhu, Xiaohua Hu, and Ming Zhang. Dialogue session segmentation by embedding-enhanced texttiling. In *INTERSPEECH*, pages 2706–2710, 2016.

[Sutton *et al.*, 2000] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, pages 1057–1063, 2000.

[Webber *et al.*, 2012] Bonnie Webber, Markus Egg, and Valia Kordoni. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490, 2012.

[Williams *et al.*, 2017] Jason D Williams, Kavosh Asadi, and Geoffrey Zweig. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *ACL*, pages 665–677, 2017.

[Williams, 1992] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

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

[Zhao *et al.*, 2011] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *ECIR*, pages 338–349, 2011.