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

Ryuichi Takanobu¹, Minlie Huang¹*, Zhongzhou Zhao², Fenglin Li², Haiqing Chen², Xiaoyan Zhu¹, Liqiang Nie³

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 utterances¹.

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-

¹Topic 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 informativeness 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 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 hand-crafted 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 re-present 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 $X = x_1, x_2, \ldots, x_T$ where each $x_i$ is an utterance, and a topic set $C = \{c_1, c_2, \ldots, c_K\}$, 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.

2The 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.
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

$$l(x) = \arg \max_j \sum_i tf(k_{ij}, x),$$

(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 $X = x_1, x_2, \ldots, x_L$, the sentence-level LSTM obtains the hidden states $H = \{h_1, h_2, \ldots, h_L\}$ 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., $h_{t-1}$ and $h_t$. 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, $\pi_\theta(a_t|S_t)$. 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., $A = \{c_1, c_2, \ldots, c_K\}$. 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 $S_t$ represents the current state after the actions $a_1, a_2, \ldots, a_{t-1}$ are sampled for the previous utterances in a dialogue session. The state can be formally represented by

$$S_t = [h_t; v(a_{t-1}); \omega_k],$$

(2)

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

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.
Policy
A stochastic policy \( \pi_\theta \) is adopted to sample a topic label for the current utterance at the current state \( S_t \). The policy function is defined by a softmax function, which is a probability distribution over all topic labels in \( C (A = C) \). 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 [Gut et al., 2017].

\[
\pi_\theta(a_t | S_t) = \text{softmax}(z; \beta) = \frac{\exp(z \cdot \beta)}{\sum_{z} \exp(z \cdot \beta)},
\]

where \( z = W \cdot S_t \) is the input to the softmax function. Obviously, a very large \( \beta \) leads to the effect of \( \arg\max \) 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

\[
r_{\text{int}} = \frac{1}{L-1} \text{sign}(a_{t-1} = a_t) \cos(h_{t-1}, h_t),
\]

where \( \text{sign}(c) \) 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 \( (h_{t-1}, h_t) \) 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:

\[
r_{\text{delayed}} = \frac{1}{N} \sum_{\omega \in \mathcal{X}} \left[ \frac{1}{|\omega|} \sum_{X_i \in \omega} \cos(h_t, \omega) - \frac{1}{N-1} \sum_{(\omega_{k-1}, \omega_k) \in \mathcal{X}} \cos(\omega_{k-1}, \omega_k), \right]
\]

where \( N \) is the number of segments predicted by the policy in a dialogue session \( X \), \( \omega \) is a topic segment in \( X \), and \( \omega \) 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 \( \tau \) as a sequence \( S_1, a_1, \cdots, S_L, a_L \)

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:

\[
J(X; \Theta) = \mathbb{E}_{(S_t, a_t) \sim P_\theta(S_t, a_t)} \left[ \sum_{t=1}^L r(S_t, a_t) \right] = \sum_{\tau \in \mathcal{T}} p(S_1) \prod_{t=1}^L \pi_\theta(a_t | S_t) p(S_{t+1} | S_t, a_t) R_L,
\]

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

\[
\nabla_\Theta J(X; \Theta) = \mathbb{E}_{\tau \sim \pi_\theta(\tau)} \left[ \sum_{t=1}^L (R_L - b(\tau)) \nabla_\Theta \log \pi_\theta(a_t | S_t) \right].
\]

(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 \( \tau \) 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.

### 3.4 Overall Procedure

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

4.1 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.

Table 2: Statistics of the corpus.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Smartphone</th>
<th>Clothing</th>
</tr>
</thead>
<tbody>
<tr>
<td># Topic category</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td># Training session</td>
<td>12,315</td>
<td>10,000</td>
</tr>
<tr>
<td># Training utterance</td>
<td>430,462</td>
<td>338,534</td>
</tr>
<tr>
<td># Gold-standard session</td>
<td>300</td>
<td>315</td>
</tr>
<tr>
<td># Gold-standard utterance</td>
<td>10,888</td>
<td>10,962</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Category</th>
<th>Keywords list</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service</td>
<td>分期(installment), 发票(receipt), 优惠(quote), 电子(electronic)</td>
</tr>
<tr>
<td>Promotion</td>
<td>退(refund), 活动(promotion), 礼(physical gift), 赠品(free gift), 分享(share)</td>
</tr>
<tr>
<td>Chitchatting</td>
<td>嗯(yeah), 谢谢(thanks), 你好(hello), 您好(yep), 客气(welcome)</td>
</tr>
<tr>
<td>Product Info</td>
<td>卡(card), 账(film), 头条(headset), 账单(model), 充电(charge)</td>
</tr>
<tr>
<td>Refunds</td>
<td>退款(refund), 退(return), 退货(fee), 同意(agree), 等回(send back)</td>
</tr>
<tr>
<td>Logistics</td>
<td>发货(delivery), 快递(express), 发(send), 达到(deliver), 预计(estimated)</td>
</tr>
<tr>
<td>Payment</td>
<td>改(change), 改变(adjust), 错(wrong), 留下(record), 无效(invalid)</td>
</tr>
</tbody>
</table>

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 $R_V = \sum_{X \in V} f(X; \theta)$ on the validation set $V$. These hyper-parameters include: the dimension of the topic vectors = 100, the learning rate $\alpha = 1e-5$ and $\beta = 0.5$ 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 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Smartphone</th>
<th>Clothing</th>
</tr>
</thead>
<tbody>
<tr>
<td>TextTiling(TeT)</td>
<td>13.09</td>
<td>.802</td>
</tr>
<tr>
<td>TopicTiling</td>
<td>3.30</td>
<td>.522</td>
</tr>
<tr>
<td>TeT+Embedding</td>
<td>3.59</td>
<td>.564</td>
</tr>
<tr>
<td>STM</td>
<td>4.37</td>
<td>.505</td>
</tr>
<tr>
<td>NL+HLSTM</td>
<td>8.25</td>
<td>.632</td>
</tr>
<tr>
<td><strong>Our method</strong></td>
<td><strong>2.69</strong></td>
<td><strong>.415</strong></td>
</tr>
</tbody>
</table>

(b) Topic Labeling (Accuracy)

<table>
<thead>
<tr>
<th>Model</th>
<th>Smartphone</th>
<th>Clothing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter-LDA</td>
<td>25.4</td>
<td>24.5</td>
</tr>
<tr>
<td>TopicTiling</td>
<td>27.6</td>
<td>17.2</td>
</tr>
<tr>
<td>Keyword Matching</td>
<td>39.8</td>
<td>31.8</td>
</tr>
<tr>
<td>NL</td>
<td>51.4</td>
<td>39.0</td>
</tr>
<tr>
<td>NL+HLSTM</td>
<td>52.6</td>
<td>40.1</td>
</tr>
<tr>
<td><strong>Our method</strong></td>
<td><strong>62.2</strong></td>
<td><strong>48.0</strong></td>
</tr>
</tbody>
</table>

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

4.4 Main Results

4.4.1 Topic Segmentation

We compared our method with the following baselines:

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

### 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
 Reinforcement Learning

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 $\Delta(A; A') = \{x | y \neq y', (x, y), (x, y') \in A \land A' \}$ 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 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


