Lexical Acquisition through Implicit Confirmation over Multiple Dialogues

Kohei Ono¹, Ryu Takeda¹, Eric Nichols², Mikio Nakano², and Kazunori Komatani¹

¹ The Institute of Scientific and Industrial Research (ISIR), Osaka University
² Honda Research Institute Japan Co., Ltd.
Our goal

• Systems that learn through dialogue
  • Acquire and accumulate knowledge during dialogues

  The Thai restaurant near the station is famous for fried rice. I like the fried rice at the Thai restaurant near the station.

• Building a closed-domain chatbot
  – Knowledge base is necessary
    • Assuming dialogue corpus including all lexical items is unrealistic
Requirements

• Chatbot
  – acquires new concepts from dialogues
    • Reduce cost to manually expand knowledge base
    • Current target: acquiring food names and categories
  – can continue dialogues even for unknown term
    • Repeating simple and abrupt questions is bothersome

2017/8/30
Our approach

- **Lexical acquisition through implicit confirmation**
  - Trying to acquire the ontological categories of unknown terms through confirmation

1. Predict category of unknown term [Otsuka et al., 13]
2. Generate implicit confirmation request with category $c$
3. Determine if the category $c$ is correct from user response
Why implicit confirmation?

• Explicit confirmation requests can be annoying
  – Category prediction often fails
  – Repeating explicit questions degrades user experience

Explicit, incorrect
Mutton Biryani was good.
Is Mutton Biryani Italian?

Explicit, too obvious
I sometimes have mushroom ravioli.
Is mushroom ravioli Italian?
Problems to solve in this work (1/2)

- Determination is difficult
  - Users respond with various expressions
  - Not only simple affirmative or negative responses

**Implicit, incorrect (easy)**

I baked Pandoro yesterday.

I want to eat Japanese food.

or

What are you talking about?

**Implicit, incorrect (not easy)**

I baked Pandoro yesterday.

I want to eat Japanese food.

I like it, too.

1. Take various features into consideration
Problems to solve in this work (2/2)

- Users do not always respond correctly
  - Their answers are sometimes inconsistent

Implicit, incorrect

I ate Mutton Biryani at a restaurant.

Japanese food is healthy, isn’t it?

Yeah.

2. Exploit responses over multiple dialogues

1. Predict category
2. Generate implicit confirmation
3. Determine correctness by user response
Our proposed methods

1. Design a feature set for ML-based classification
   – Considering expressions other than affirmative/negative ones
   – Exploiting user utterances around implicit confirmation request

2. Exploit determination results over multiple dialogues
   – Possible when the system is deployed on a server
Overview of our methods

Method #1
Calculate probability $p_i(w, c)$ from single user response $i$

Obtain $n$ user responses

Method #2
Calculate $Conf(w, c)$ from $p_1(w, c), \ldots, p_n(w, c)$

$Conf(w, c)$
Today’s agenda

1. Background and our approach
2. Proposed methods
   – Determine if a category in the user response is correct
3. Method #1: Feature set for single user responses
   – Data collection
   – Experiment: Performance for single user responses
4. Method #2: Exploiting responses over multiple dialogues
   – Experiment: Improvement by using multiple user responses
5. Conclusion and future work
Method #1: Feature set for single user responses

Method #1

Calculate probability $p_i(w, c)$ from single user response $i$

Obtain $n$ user responses

Method #2

Calculate $Conf(w, c)$ from $p_1(w, c), \ldots, p_n(w, c)$
Calculate probability $p_i(w, c)$ with a single user response

• Whether category $c$ for unknown term $w$ in confirmation request is correct or not

• Logistic regression (LR) with 11 features:
  a. Expressions in U2 (not only affirmative/negative)
  b. U1 expressions and its relationship with S1
  c. Relationship between U1 and U2
Designed feature set

- f1: U2 includes an expression affirmative to S1
- f2: U2 includes an expression negative to S1
- f3: U2 includes an expression correcting S1
- f4: U2 includes the category name used in S1
- f5: U2 includes a category name not used in S1, excluding cases that fall under f3
- f6: U2 includes a word preventing topic change in S1
- f7: U1 includes the category name used in S1
- f8: U1 includes a category name not used in S1
- f9: U1 includes any interrogative
- f10: U1 includes an expression corresponding to the category mentioned in S1
- f11: U1 and U2 contain the same word

All features are binary
Data collection via crowdsourcing

- Collect user utterances before/after implicit confirmation requests
  
  Specified term
  
  I ate *bagna cauda* for the first time.

  Correct
  
  Italian is perfect for a date, isn’t it? or
  
  I love a *meat dish*.

  Incorrect

- 20 terms and their corresponding implicit confirmation requests
- 1,956 dialogues (U1-S1-U2) from 98 workers
Results – Experiment #1

- LR, 10-fold cross validation
- Category $c$ is regarded as correct if $p_i(w, c) \geq 0.5$

<table>
<thead>
<tr>
<th>Features</th>
<th>Output</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1, f2</td>
<td>Correct</td>
<td>320</td>
<td>220</td>
<td>0.593</td>
</tr>
<tr>
<td>(Baseline)</td>
<td>Incorrect</td>
<td>658</td>
<td>758</td>
<td>0.535</td>
</tr>
<tr>
<td>All</td>
<td>Correct</td>
<td>742</td>
<td>313</td>
<td>0.703</td>
</tr>
<tr>
<td>(Proposed)</td>
<td>Incorrect</td>
<td>236</td>
<td>665</td>
<td>0.738</td>
</tr>
</tbody>
</table>

Baseline: only uses affirmative and negative expressions

Significant feature: $f_7$ (U1 includes the category name used in S1)
Insignificant feature: $f_9$ (U1 includes any interrogative)

**Proposed features improve detection of correct categories**
Today’s agenda

1. Background and our approach
2. Proposed methods and procedure
   – Determine if a category in the user response is correct
3. Method #1: Feature set for single user responses
   – Data collection
   – Experiment: Performance for single user responses
4. Method #2: Exploiting responses over multiple dialogues
   – Experiment: Improvement by using multiple user responses
5. Conclusion and future work
Proposed #2: Exploiting responses over multiple dialogues

Method #1

Calculate probability $p_i(w, c)$ from single user response $i$

Obtain $n$ user responses

Method #2

Calculate $Conf(w, c)$ from $p_1(w, c), \ldots, p_n(w, c)$

$Conf(w, c)$
Calculate $Conf(w, c)$ from multiple responses

- To determine correct categories from $n$ user responses
  - Also logistic regression (LR)
  - 5 features from $p_i(w, c)$
    
    $$(i = 1, ..., n)$$

<table>
<thead>
<tr>
<th>g1</th>
<th>Average of $p_i(w, c)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>g2</td>
<td>$n$</td>
</tr>
<tr>
<td>g3</td>
<td>$\max_i p_i(w, c)$</td>
</tr>
<tr>
<td>g4</td>
<td>$\min_i p_i(w, c)$</td>
</tr>
<tr>
<td>g5</td>
<td>${i</td>
</tr>
</tbody>
</table>

- Category $c$ is regarded as correct for unknown term $w$ when $Conf(w, c)$ exceeds a threshold
  - Then add them to system’s knowledge base
Experiment #2 -- Conditions

• Dataset
  – 1,956 responses from 98 workers
  – Divided them into training and test sets
    • Perfectly open with workers and requests
  – Selected $n$ responses having $p_i(w, c)$ randomly from 49 (or 48) workers’ responses
    • $n$ is 2 to 49 (or 48); 1,000 combinations at most
  – Obtained the feature values from the $n$ responses and calculate $Conf(w, c)$
Experiment #2 -- Results

Questions to answer:

a. Does the regression performance improve by multiple user responses?

b. How many responses are needed to acquire reasonably correct categories?

c. How can we set a threshold for $Conf(w, c)$?
a) Does regression performance improve?

- BEP values of $n \geq 2$ were larger than that of $n = 1$
  - BEP (breakeven point): the value when precision = recall

Multiple user responses are helpful to determine if the predicted category is correct
b) How many user responses are needed?

- Increases in BEP values were large when $n \leq 5$
  - It is worthwhile to ask more users
- They diminished when $n \geq 10$
  - Diminished returns from asking more users
c) How can we set the threshold?

- High precision rates are required
  - To avoid acquiring incorrect information
- Set high thresholds so that precision rate becomes almost one (0.995)
  - The recall rate for \( n = 5 \) was 0.175
  - Recall rates gradually increased with \( n \)

Sufficiently-high thresholds enable the system to acquire more categories of unknown terms with high precision rate
Summary and future work

Towards systems that learn through dialogues

• Goal in this paper
  – Determine if the ontological category of an unknown term included in an implicit confirmation request is correct or not

• Proposed methods
  1. Design a feature set for single user responses
  2. Calculate $\text{Conf}(c, w)$ over multiple user responses

• Results
  – Performance improved by (1) response context and (2) multiple user responses

• Future work (on-going)
  – User study to compare implicit and explicit confirmation requests
  – Incorporating the proposed method into a deployed chatbot
以下、予備スライド
関連研究

- 質問を利用した未知語獲得
  - 単純な質問[菅生, 14]
  - 未知語の概念に関する質問[大塚, 13]

- 明示的確認要求

システムが何度も質問を行うとユーザは煩わしく感じる
あるクラスcについての単一ユーザのやりとりから確信度\( p_i(c) \)算出

ステップ1を\( n \)人に対して行う

提案手法②

\( p_1(c), ..., p_i(c), ..., p_n(c) \)から確信度\( Conf(c) \)を算出

提案手法①
確認要求への同意表現の有無

• 正しい確認要求に対してユーザは同意する傾向

2017/8/30
確認要求前後での共通した単語

• 正しいクラスを含む確認要求に対し、ユーザはU1で用いた単語についての話をU2でも紹続する

東南アジアの国々でカオマンガイは食べられるよ

最近では日本でも手軽にエスニックフードが食べられますね

日本でもカオマンガイのおいしいお店があるよね

– f11で表現

2017/8/30
話題転換の防止

・誤った確認要求によって逸れた話題をユーザは元に戻そうとする

サングリアのフルーティーな甘さが好き
洋菓子は味が濃厚なものが多いですよね
お酒の話だよ？

– U2中に、
  1. 確認要求に用いていないクラス（f5で表現、全25クラス）
  2. 話題転換を防ぐための「話」という語（f6で表現）
    が出現
### 素性への重み(単一)

<table>
<thead>
<tr>
<th>素性</th>
<th>係数</th>
</tr>
</thead>
<tbody>
<tr>
<td>確認要求への同意表現の有無</td>
<td>0.268</td>
</tr>
<tr>
<td>確認要求への反対表現の有無</td>
<td>-3.281</td>
</tr>
<tr>
<td>確認要求への訂正表現の有無</td>
<td>-4.390</td>
</tr>
<tr>
<td>確認要求に用いられたクラスの有無</td>
<td>-1.076</td>
</tr>
<tr>
<td>確認要求に用いられていないクラスの有無</td>
<td>-0.867</td>
</tr>
<tr>
<td>確認要求による話題転換を防止する語の有無</td>
<td>-1.508</td>
</tr>
<tr>
<td>確認要求に用いられたクラスの有無</td>
<td>-2.645</td>
</tr>
<tr>
<td>確認要求に用いられていないクラスの有無</td>
<td>35.358</td>
</tr>
<tr>
<td>疑問詞の有無</td>
<td>-1.415</td>
</tr>
<tr>
<td>U1中の表現と確認要求に用いられたクラスとの一致</td>
<td>-0.221</td>
</tr>
<tr>
<td>U1とU2とで共通した単語の有無</td>
<td>-4.213</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.774</td>
</tr>
</tbody>
</table>

確認要求前後のユーザ発話(U1)の表現

確認要求前後のユーザ発話間の関係
素性への重み（複数）

<table>
<thead>
<tr>
<th></th>
<th>グループ1</th>
<th>グループ2</th>
<th>グループ3</th>
<th>グループ4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_i(c)$の平均値</td>
<td>43.5797</td>
<td>63.7355</td>
<td>38.2668</td>
<td>4.9976</td>
</tr>
<tr>
<td>$n$の値</td>
<td>-0.0067</td>
<td>-0.0096</td>
<td>-0.0004</td>
<td>0.0397</td>
</tr>
<tr>
<td>$p_i(c)$の最大値</td>
<td>6.1139</td>
<td>5.7463</td>
<td>12.0324</td>
<td>9.7411</td>
</tr>
<tr>
<td>$p_i(c)$の最小値</td>
<td>-1.3051</td>
<td>-5.3844</td>
<td>3.9986</td>
<td>14.5041</td>
</tr>
<tr>
<td>$(p_i(c) \geq 0.5$となる応答数) / n</td>
<td>-2.9270</td>
<td>-11.6434</td>
<td>-2.6292</td>
<td>3.4047</td>
</tr>
<tr>
<td>Intercept</td>
<td>-25.4867</td>
<td>-29.3268</td>
<td>-29.5732</td>
<td>-15.8032</td>
</tr>
</tbody>
</table>
データセットの作成

・1956のやりとりから得られた$p_i(c)$を、ワークと確認要求との観点から4つのグループに分割
  - ユーザopen、確認要求openの評価を行うため

<table>
<thead>
<tr>
<th></th>
<th>ワーク #1〜#49</th>
<th>ワーク #50〜#98</th>
</tr>
</thead>
<tbody>
<tr>
<td>正しい確認要求 #1〜#5</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>誤った確認要求 #1〜#5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>正しい確認要求 #6〜#10</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>誤った確認要求 #6〜#10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>