Incorporating Discourse Features into Confidence Scoring of Intention Recognition Results in Spoken Dialogue Systems

Ryuichiro Higashinaka, Katsuhito Sudoh, and Mikio Nakano
NTT Communication Science Laboratories
Overview

• A new confidence scoring method for intention recognition results in spoken dialogue systems
  – Intention means the information that the user wants to convey to the system
  – Uses *discourse features* in addition to acoustic and language model features
  – Useful for dialogue management e.g., avoid unnecessary confirmations
### Intention Recognition: an example

<table>
<thead>
<tr>
<th>Frame 1</th>
<th>Place</th>
<th>Date</th>
<th>Info</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

#### Example Dialogue

**System**: “May I help you?”

**User**: “Tell me *Tokyo’s* weather for tomorrow”

(Tokyo was misrecognized as *Kyoto*)

**System**: “Kyoto’s weather for tomorrow?”

**User**: “Tokyo”

(Tokyo was misrecognized as *Kyoto again*)

**System**: “Kyoto’s weather for tomorrow?”

**User**: “No, *Tokyo!*”

---

<table>
<thead>
<tr>
<th>Frame 2</th>
<th>Place</th>
<th>Date</th>
<th>Info</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kyoto</td>
<td>tomorrow</td>
<td>weather</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frame 3</th>
<th>Place</th>
<th>Date</th>
<th>Info</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kyoto</td>
<td>tomorrow</td>
<td>weather</td>
</tr>
</tbody>
</table>

**Confidence = ?**

<table>
<thead>
<tr>
<th>Frame 4</th>
<th>Place</th>
<th>Date</th>
<th>Info</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tokyo</td>
<td>tomorrow</td>
<td>weather</td>
</tr>
</tbody>
</table>

---
Conventional Methods

Use **confidence of words** in speech recognition results for the confidence of slot values.

**Frame2**

<table>
<thead>
<tr>
<th>Place</th>
<th>Kyoto</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>tomorrow</td>
</tr>
<tr>
<td>Info</td>
<td>weather</td>
</tr>
</tbody>
</table>

User: “Tell me Kyoto’s weather for tomorrow.”

System: “Kyoto’s weather for tomorrow?”

User: “Kyoto”

System: “Kyoto’s weather for tomorrow?”

User: “No, Tokyo!”

**Frame3**

<table>
<thead>
<tr>
<th>Place</th>
<th>Kyoto</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>tomorrow</td>
</tr>
<tr>
<td>Info</td>
<td>weather</td>
</tr>
</tbody>
</table>

**Frame4**

<table>
<thead>
<tr>
<th>Place</th>
<th>Tokyo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>tomorrow</td>
</tr>
<tr>
<td>Info</td>
<td>weather</td>
</tr>
</tbody>
</table>

Use confidence of words in speech recognition results for the confidence of slot values.
Proposed Method

• Slot value is not a result of a single utterance but the entire discourse
  - Use discourse information to improve accuracy of confidence scoring

• **Train a confidence model** that outputs confidence scores based on both
  - acoustic and language model features of a word filling the slot and
  - *discourse features* for the slot value
Discourse Features

System: “May I help you?”
User: “Tell me **Tokyo**’s weather for tomorrow”
System: “Kyoto’s weather for tomorrow?”
User: “**Tokyo** (Kyoto)”
System: “Kyoto’s weather for tomorrow?”
User: “No, **Tokyo**!”

**Discourse features encode the relationship between a slot value and the discourse**
Discourse Features (cont’d)

• We enumerated 11 discourse features
  – How many times the same slot value is found in previous frames
  – Ratio of the slot value in all frames
  – How many times the slot value was deleted or overwritten by other values
  – How many times the slot value has appeared in user and system utterances
  – etc.
Discourse Features (cont’d)

• Same keyword pair count
  – The number of times the slot value is confirmed by the system and then uttered by the user immediately afterwards
  – System : “Kyoto’s weather for tomorrow?”
    User : “Kyoto”
  – Grice’s maxim of quantity states that user utterances have to be as informative as necessary
  – Possible penalty to slot values that are related to this less informative interaction
Data Collection

• System
  – Weather Information Service Domain
  – Vocabulary of 1,652 words
  – Has 3 slots (place, date, information-type)

• Collected data
  – 18 subjects performed 16 dialogues each
  – 288 dialogues collected
  – Task completion rate is 95.83% (276/288)
  – 4812 slot value samples
Data Screening

- Slots that did not have values
- Slots explicitly confirmed by the user
- Slots that have only one value in all frames

Kyoto and weather have the same discourse features although one of them is wrong.

It causes trouble in confidence model training.

777 slot samples remained.
Confidence Model Training

- **Feature extraction**
  - 27 acoustic and language model features adopted from *(Hazen et al. 2002)*
  - 11 discourse features

- **Confidence model**
  - Weighted linear combination of the features adopted from *(Hazen et al. 2002)*
  - Weights are optimized using the training data
  - Outputs positive scores for correct slot values and negative scores for incorrect ones
Evaluation

• Comparison of two confidence models
  – Conventional Model (conv.)
    • trained only by acoustic and language model features
  – Proposed Model (prop.)
    • trained by both acoustic and language model features and discourse features

• 6-fold cross validation
Evaluation (cont’d)

Proposed model outperforms conventional model in classification accuracy
Evaluation (cont’d)

- Matrix of counts of correct and incorrect items

<table>
<thead>
<tr>
<th></th>
<th>Prop. Correct</th>
<th>Prop. Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv. Correct</td>
<td>535</td>
<td>35</td>
</tr>
<tr>
<td>Conv. Incorrect</td>
<td>102</td>
<td>105</td>
</tr>
</tbody>
</table>

Statistically significant difference in classification performance (McNemar’s test, $p = 8.69 \cdot 10^{-8}$)
Impact of the discourse features

• relatively important features
  – Same keyword pair count
    □ Slot values related to the less informative interaction is likely to be incorrect
  – Number of slot values in user utterance
    □ The more the slot value is found in user utterances, the more correct the slot value is

• less important feature
  – Ratio of the slot value in all frames
    □ Ratio in frames does not guarantee its correctness
Conclusion

• A new confidence scoring method for intention recognition results in spoken dialogue systems
  – Uses discourse features in addition to acoustic and language model features
• Experimental Results show validity of our method
• *Future work:*
  – Verification in other domains
  – Online evaluation of the system