ALADIN

The development of self-learning assistive vocal interfaces for people with physical impairments

Dag van de Fonetiek, 15 December 2011

Janneke van de Loo, Guy De Pauw, Jort Gemmeke, Peter Karsmakers, Bert van den Broeck, Walter Daelemans, Hugo Van hamme
ALADIN - introduction

Demand from health care sector:

**Vocal user interfaces** for people with **physical impairments**, for controlling devices at home (TV, radio, lights, etc.)
ALADIN - introduction

Problems with existing vocal interfaces:

• high development costs because of large variation in user needs
• lack of robustness of speech recognizer to:
  ▪ user-specific speech characteristics (pathological speech, regional pronunciation variation)
  ▪ environmental noise
• predefined vocabulary & grammar → learning and adaptation required from user
ALADIN - introduction

→ **ALADIN**: Adaptation and Learning for Assistive Domestic Vocal Interfaces

**Goal**: develop a robust, self-learning domestic vocal interface that adapts to the user instead of the other way around:
- learn the user’s vocabulary & grammar constructs
- learn the user’s voice & pronunciation characteristics

**How?** Unsupervised learning on the basis of training examples: vocal commands + associated controls (actions)
ALADIN - introduction

Target applications:
- control devices such as TV, lights, doors, heating, etc.
- play games, e.g. the card game *patience* (*solitaire*)
ALADIN - grammar induction

Grammar: maps the structure of the sentence to its meaning – in this case: an action

command  grammar  meaning (action)
ALADIN - grammar induction

Grammar: maps the structure of the sentence to its meaning – in this case: an action

command \[\rightarrow\] grammar \[\rightarrow\] meaning (action)

‘Zet de TV iets luider’
*(Turn the TV a bit louder)*

‘Open de badkamerdeur’
*(Open the bathroom door)*

frame-based representation
ALADIN - grammar induction

Grammar: maps the structure of the sentence to its meaning – in this case: an action

command → grammar → meaning (action)

‘Zet de TV iets luider’
(Turn the TV a bit louder)

‘Open de badkamerdeur’
(Open the bathroom door)

‘Leg de harten drie op de klaveren vier’
(Put the three of hearts on the four of clubs)

‘Leg de harten vier in kolom drie’
(Put the four of hearts in column three)
Patience data collection

- 8 participants, playing patience using voice commands
- commands are executed by the experimenter
- half of the participants: Wizard-of-Oz
- each participant 2 x 30 mins (with at least 3 weeks in between)

Participants:
- 4 men, 4 women
- 4 higher educated, 4 lower educated
- ages between 23 and 73
Patience data collection

Resulting files per move:
• Recorded: commands (audio)
• Generated by Patience program (Matlab):
  - Frame description of action
  - Complete workspace (state of the game)

The commands are orthographically transcribed and annotated.
Patience data: example

situation on screen

command:
‘de ruiten vier op de schoppen vijf’
(the four of diamonds on the five of spades)
Patience data: example

situation on screen

command:
‘de ruiten vier op de schoppen vijf’
(the four of diamonds on the five of spades)
### Action frame: ‘MoveCard’

<table>
<thead>
<tr>
<th>Frameslot</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>From_Suit</td>
<td></td>
</tr>
<tr>
<td>From_Value</td>
<td></td>
</tr>
<tr>
<td>From_Foundation</td>
<td></td>
</tr>
<tr>
<td>From_FieldCol</td>
<td></td>
</tr>
<tr>
<td>From_FieldRowFaceUp</td>
<td></td>
</tr>
<tr>
<td>From_FieldRowAll</td>
<td></td>
</tr>
<tr>
<td>From_Hand</td>
<td></td>
</tr>
<tr>
<td>To_Suit</td>
<td>Spades</td>
</tr>
<tr>
<td>To_Value</td>
<td></td>
</tr>
<tr>
<td>To_Foundation</td>
<td></td>
</tr>
<tr>
<td>To_FoundationEmpty</td>
<td></td>
</tr>
<tr>
<td>To_FieldCol</td>
<td></td>
</tr>
<tr>
<td>To_FieldColEmpty</td>
<td></td>
</tr>
<tr>
<td>To_FieldRowFaceUp</td>
<td></td>
</tr>
<tr>
<td>To_FieldRowAll</td>
<td></td>
</tr>
<tr>
<td>HorizontalMovement_Distance</td>
<td>1</td>
</tr>
<tr>
<td>HorizontalMovement_Direction</td>
<td>Left</td>
</tr>
</tbody>
</table>

Diamonds 4-612-  
Spades 5--5-24  
‘de ruiten vier op de schoppen vijf’
Grammar Induction

Sequence tagging / chunking approach:
Identify words, or chunks of words, referring to specific frame slots

\[
\begin{align*}
\text{de ruiten} & \quad \text{vier} & \quad \text{op} & \quad \text{de} & \quad \text{schoppen} & \quad \text{vijf} \\
\text{O} & \quad \text{FromSuit} & \quad \text{FromValue} & \quad \text{O} & \quad \text{O} & \quad \text{ToSuit} & \quad \text{ToValue}
\end{align*}
\]

\[
\begin{align*}
\text{de schoppen} & \quad \text{koning} & \quad \text{op} & \quad \text{de} & \quad \text{lege} & \quad \text{plaats} \\
\text{O} & \quad \text{FromSuit} & \quad \text{FromValue} & \quad \text{O} & \quad \text{O} & \quad \text{ToColEmpty} & \quad \text{ToColEmpty}
\end{align*}
\]

\[
\begin{align*}
\text{de harten} & \quad \text{drie} & \quad \text{afleggen} & \quad \text{bovenaan} \\
\text{O} & \quad \text{FromSuit} & \quad \text{FromValue} & \quad \text{ToFoundation} & \quad \text{ToFoundation}
\end{align*}
\]
Grammar Induction

Initially: experiments with text data (transcriptions)

Supervised grammar induction (tagging):
→ training data: commands annotated with “frame slot tags”

```
de ruiten    vier    op    de    schoppen    vijf
O  FromSuit  FromValue  O  O  ToSuit  ToValue
```

→ task: assign “frame slot tags” to commands in test set

Later: Unsupervised grammar induction
Supervised grammar induction

Preliminary experiments

• Data:
  - participants 1–4
  - around 270 utterances per participant except participant #2: 171 utterances

• Division train/test (per participant):
  - training data: first $n$ utterances
  - test data: the rest

• Tool: MBT (Memory-Based Tagger Generator and Tagger$^1$)

$^1$ http://ilk.uvt.nl/mbt/
### Supervised grammar induction

Average results participants 1-4:

<table>
<thead>
<tr>
<th>Train size (#utterances)</th>
<th>Tag Accuracy(%)</th>
<th>Known words</th>
<th>Unknown words</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>93.9</td>
<td>53.7</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>96.1</td>
<td>53.4</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>97.7</td>
<td>29.3</td>
<td></td>
</tr>
</tbody>
</table>
Future work

- Supervised experiments:
  - learning curves
  - add 2\textsuperscript{nd} layer of tags referring to frame slot values, experiment with shared learning of values

- Unsupervised experiments
  - initially with text as input
  - later with hypotheses from word finding module as input
Questions?