A DISCRIMINATIVE APPROACH TO GROUNDED SPOKEN LANGUAGE UNDERSTANDING IN INTERACTIVE ROBOTICS

EMANUELE BASTIANELLI (*), DANilo CROCE (§), ANDREA VANZO (*), ROBERTO BASILI (§), DANIELE NARDI (*)

(*) University of Roma, La Sapienza
(§) University of Roma, Tor Vergata

IJCAI 2016, New York, July 12th, 2016
OVERVIEW

• Motivations
• The SLU chain setting
• The SLU semantic tasks
  • Learning Models
  • Perceptual Features
• Experimental Results
• Understanding human language is a crucial problem for the Robotics area as much as the systems are expected to enter our daily life

• Differences with traditional SLU:
  • Robots acts in the real world, and the language they receive is grounded, i.e. include linguistic references to the real world that must be explicited by the interpretation
"take the book on the table"

**TAKE action**
1) go to the table
2) take the book

**BRING action**
1) go to the shelf
2) take the book
3) go to the table
4) release the book

One-step process:
1) go to the table
2) take the book

Two-step process:
1) go to the shelf
2) take the book
3) go to the table
4) release the book

---

**GROUNDED SLU**
1. Injection of *perceptual information* directly in the *language learning* algorithm for effective SLU
2. Integration of distributional lexical semantic information
3. Use of sound *linguistic theories*, i.e. Framenet
4. Exploiting a large *corpus* for training and benchmarking
• What does perceptual SLU mean from a Machine Learning perspective?

• Design of grounding properties able to describe actual environmental settings

• Model properties as features directly injected in the learning process

• Design different features for different semantic stages of the SLU chain

\[ f_1: \text{there is a book} \]
\[ f_2: \text{there is a table} \]
\[ f_3: \text{the book is on the table} \]
\[ f_4: \text{the book is near the table} \]

...
NATURAL LANGUAGE, WORLD REPRESENTATION AND ROBOTS

Language Level

Platform Level

Semantic Map

Real World

Frame: Taking

THEME: the book on the table

plan: Grab

:hasArg

argument: Object

table

book

tl

bl

Platform Model

take the book on the table
THE SLU PROCESSING PIPELINE

Automatic Speech Recognition
Morpho-Syntactic Analysis
ASR Re-ranking
Semantic Parsing
Grounding

“take the book”

dig dig book
take the book
take the Booth
dig dig book

[take] Taking
[the book] Theme

grab(obj:b012, pos:[10,3])
SEMANTIC PARSING AS A CLASSIFICATION STAGE

• Provide an interpretation of an utterance in terms of Frame Semantics

  *Take the book on the table*

• Performed in three steps

  1. **Action Detection**
     
     \[\text{take}\]_{\text{Taking}} \text{the book on the table}

  2. **Argument Identification**
     
     \[\text{take}\]_{\text{Taking}} \text{[the book on the table]}

  3. **Argument Classification**
     
     \[\text{take}\]_{\text{Taking}} \text{[the book on the table]}_{\text{THEME}}
SEMANTIC PARSING AS SEQUENCE LABELING

• Sequence Labeling with SVM-HMM (Altun et al., 2003)
• Hybrid discriminative-generative approach
SLU PROCESSING PIPELINE

“take the book”

Automatic Speech Recognition
Morpho-Syntactic Analysis
ASR Re-ranking
Semantic Parsing
Grounding

dig dig book
take the book
take the Booth

[take]_taking
[the book]_theme

grab(obj:b012,
pos:[10,3])
LEXICALIZED GROUNDING FUNCTION

• Semantic Maps “say” a lot about the environment
  
  • E.g.
    - There is an entity called *table*
    - There is an entity called *book*
    - The *book* and the *table* are in some spatial relation. In this case, they are generally *close*.

• Find a way to express such information *numerically*
LEXICALIZED GROUNDING FUNCTION (2)

• Lexicalized Grounding Function $LG(w, SM)$
  • Abductive mechanism that expresses a confidence score about the grounding of a word $w$ against a semantic map $SM$
  • It jointly uses of *Distributional Models of Lexical Semantics* and *Edit Distances* over transcribed words ($w$) and entity names ($w_e$) in $SM$

1. Being $S^w$ a set of word semantically related to $w$ in the DM
2. Being $sim(.,.)$ a similarity function between word vectors (in DM)
3. Being $ph(.,.)$ a phonetic distance metric

$$LG(w, SM) = \{ e \in SM \mid \max_{w_d \in S^w} (ph(w, wd) \cdot sim(w_e, wd) + sim(w, w_e)^2) > \tau \}$$
LEXICALIZED GROUNDING FUNCTION (3)

\[ \mathcal{LG}(w, SM) = \{ e \in SM \mid \max_{w_d \in S^w} (ph(w, w_d) \cdot sim(w_e, w_d) + sim(w, w_e)^2) > \tau \} \]
THE EVALUATION CORPUS: HURIC

• **Human-Robot Interaction Corpus**\(^1\) (HuRIC), multi-layered corpus for house servicing robotics

• Data sets characterized by language complexity
  – Grammar Generated (GG) – 137/48
  – Speaky for Robots (S4R) – 145/95
  – RoboCup (RC) – 288/174
  – Rockin Competition (RoCo) – 179/120
  – Rockin Camp (RoCa) – 134/114

• 16 Frames for about 500 examples, 1000 roles

• Gathered mostly in the context of the RoCKIn Project\(^2\)

\(^2\) [http://rockinrobotchallenge.eu/](http://rockinrobotchallenge.eu/)
PERCEPTUAL FEATURES AND SEMANTIC PARSING

\[
\text{[take]}_{\text{Taking}} \quad \text{[the book on the table]}_{\text{THEME}}
\]

- Only adopting linguistic features for training

<table>
<thead>
<tr>
<th></th>
<th>Action Detection</th>
<th>Argument Identification</th>
<th>Argument Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(F1)</td>
<td>(RER)</td>
<td>(F1)</td>
</tr>
<tr>
<td>noPM/noSM</td>
<td>94.6</td>
<td>0.0</td>
<td>noSM</td>
</tr>
</tbody>
</table>
Semantic Parsing: Perceptual Features

Linguistic Features

Perceptual Features
[take] the book on the table

avg. dist. = 0.04 (close)

avg. dist. = 3.31 (far)
SEMANTIC PARSING: ARGUMENT IDENTIFICATION

\[
\text{[take]}_{Bringing} \quad \text{on} \quad \text{the} \quad \text{table}
\]

\[
\text{[take]}_{Taking} \quad \text{[the book] on the table}
\]

\[
\text{[take]}_{Bringing} \quad \text{[the book] on the table}
\]

avg. dist. = 0.04 (near)

avg. dist. = 3.31 (far)
SEMANTIC PARSING: ARGUMENT CLASSIFICATION

• During **Argument Classification**, spatial features appear less relevant
• The semantics of argument heads is much more relevant
  • It characterizes selectional restrictions expressed through the classes and attributes of entities
• Distributional Models of lexical semantics for smoothing

\[
\text{[take]}_{\text{Taking}} \left[ \text{the book on the table} \right]_{\text{THEME}}
\]
\[
\text{[grab]}_{\text{Taking}} \left[ \text{the volume on the table} \right]_{\text{THEME}}
\]
SEMsANTIC PARSING RESULTS

• Results on the individual stand-alone tasks (F1, RER)

<table>
<thead>
<tr>
<th>Action Detection</th>
<th>F1</th>
<th>RER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>73.7</td>
<td>-</td>
</tr>
<tr>
<td>noPM/noSM</td>
<td>94.6</td>
<td>0.0</td>
</tr>
<tr>
<td>onlyPM</td>
<td>95.1</td>
<td>7.9</td>
</tr>
<tr>
<td>Perc-GG</td>
<td>96.3</td>
<td>31.1</td>
</tr>
<tr>
<td>Perc-\textbf{LG}</td>
<td>95.9</td>
<td>23.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Argument Identification</th>
<th>F1</th>
<th>RER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>83.4</td>
<td>-</td>
</tr>
<tr>
<td>noSM</td>
<td>90.7</td>
<td>0.0</td>
</tr>
<tr>
<td>Perc-GG</td>
<td>94.6</td>
<td>41.9</td>
</tr>
<tr>
<td>Perc-\textbf{LG}</td>
<td>94.3</td>
<td>38.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Argument Classification</th>
<th>F1</th>
<th>RER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>21.8</td>
<td>-</td>
</tr>
<tr>
<td>noDM</td>
<td>94.3</td>
<td>0.0</td>
</tr>
<tr>
<td>DM</td>
<td>\textbf{95.3}</td>
<td>\textbf{15.8}</td>
</tr>
</tbody>
</table>
SEMANTIC PARSING RESULTS

- Results on the end-to-end process
  - Require perfect command grounding

```
<table>
<thead>
<tr>
<th>AD</th>
<th>AI</th>
<th>AC</th>
<th>Grounding</th>
<th>F1</th>
<th>F1</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>noPM/noSM</td>
<td>noSM</td>
<td>noDM</td>
<td><strong>Identity</strong></td>
<td>44.3</td>
<td>41.7</td>
<td>29.8</td>
</tr>
<tr>
<td>noPM/noSM</td>
<td>noSM</td>
<td>noDM</td>
<td><strong>LG</strong></td>
<td>68.3</td>
<td>66.7</td>
<td>50.8</td>
</tr>
<tr>
<td>Perc-<strong>LG</strong></td>
<td>noSM</td>
<td>noDM</td>
<td><strong>LG</strong></td>
<td>70.4</td>
<td>66.7</td>
<td>51.2</td>
</tr>
<tr>
<td>Perc-<strong>LG</strong></td>
<td>Perc-<strong>LG</strong></td>
<td>noDM</td>
<td><strong>LG</strong></td>
<td>73.6</td>
<td>66.4</td>
<td>51.4</td>
</tr>
<tr>
<td>Perc-<strong>LG</strong></td>
<td>Perc-<strong>LG</strong></td>
<td>DM</td>
<td><strong>LG</strong></td>
<td>74.6</td>
<td>67.3</td>
<td>52.1</td>
</tr>
</tbody>
</table>
```
CONCLUSIONS

• The proposed learning approach for grounded SLU is very effective and makes explicit and modular use of use of perceptual information directly in the language learning stage
  • It can independently accommodate more complex spatial modeling mechanisms
  • It can be applied in an on-learning fashion supporting robots evolving over time

• The reference linguistic theories provide a solid meaning representation formalism for HRI and supports extensions to a large range of linguistic phenomena

• The adopted benchmarking corpus, HuRIC, is open and available for training and benchmarking in HRI and Interactive Robotics http://sag.art.uniroma2.it/demo-software/huric

• The SLU chain corresponding to the proposed model is available for downloading at:

  http://sag.art.uniroma2.it/sluchain.html