Learning to Parse Sentences with Cross-Situational Learning using Different Word Embeddings Towards Robot Grounding

Spatial Language Understanding and Grounded Communication for Robotics Workshop, ACL-IJCNLP 2021

Subba Reddy Oota, Frédéric Alexandre, Xavier Hinaut
1INRIA Bordeaux Sud-Ouest, France
{subba-reddy.oota, frederic alexandre, xavier.hinaut}@inria.fr

Introduction

- Grounded Language Acquisition:
  1. it is the process of learning a language
  2. how children can learn language by observing their environments, interacting with others, understanding the concepts of a language (i.e., word-to-meaning) as it relates to the physical world.
- Cross-situational Learning (CSL): Understanding the mechanism enabling children to learn rapidly word-to-meaning mapping in uncertain conditions.

Why to perform grounded language acquisition through CSL?

- Acquiring language is not a supervised task: e.g. before one year of age, children can segment words from speech based on statistical learning mechanisms.
- What children observe while hearing the “the red cup is on the right” and how they map these sounds to multi-modal features, learning the concept as it refers to a red or blue cup.
- It is still not understood how meaning concepts are captured from complex sentences, along with learning language-based interactions.
- Also, how pre-trained transformer models perform grounded language acquisition through cross-situational learning (CSL) remains unclear.
- Such systems could benefit the field of human-robot interactions and help understand how children learn and ground language.

Main Contributions

- We introduce a fine-tuned BERT using the masked-language modeling objective trained on a language corpus (i.e. Juven’s CSL + GoLD).
- We showcase that One-Hot and BERT fine-tuned representations significantly improve the stimulated vision’s prediction than pre-trained Google BERT.
- We interpret the inner workings of both models and plot the evolution of the output activation during the processing of a sentence.

Approach

To build the grounded language acquisition models to employ a CSL task using two sequence models:
- Echo State Networks (i.e Reservoir Computing) – ESN
  1. ESN with Final Learning: the online algorithm is applied to the reservoir state after the last word of the sentence.
  2. ESN with Continual Learning: the reservoir states are updated after each word of a sentence using the online method
- Long Short-Term Memory Networks (LSTM)

Evaluation Metrics

- “the cap is on the right”

<table>
<thead>
<tr>
<th>Imagined vision</th>
<th>is valid?</th>
<th>is exact?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>(b)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(c)</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Dataset

- **Juven’s CSL Dataset**: It is composed of 1000 training sentences, 1000 testing sentences, where each sentence is describing one or two objects.
- **Grounded language dataset (GoLD)**: There are 8250 textual descriptions consists of 47 object classes spread across five different groups, 7 actions, and 8 colors.

Experimental Results

Baseline Results Comparison

<table>
<thead>
<tr>
<th>Model</th>
<th>Valid Error</th>
<th>Exact Error</th>
<th>Valid Error</th>
<th>Exact Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESN FL + One-Hot</td>
<td>0.28</td>
<td>1.64</td>
<td>0.05</td>
<td>49.1</td>
</tr>
<tr>
<td>ESN FL + BERT CSL</td>
<td>0.28</td>
<td>1.64</td>
<td>0.05</td>
<td>49.1</td>
</tr>
<tr>
<td>ESN CL + Google BERT</td>
<td>0.2</td>
<td>7.72</td>
<td>26.8</td>
<td>51.48</td>
</tr>
<tr>
<td>ESN CL + One-Hot</td>
<td>2.32</td>
<td>12.11</td>
<td>20.16</td>
<td>49.5</td>
</tr>
<tr>
<td>ESN CL + BERT CSL</td>
<td>2.41</td>
<td>13.7</td>
<td>18.19</td>
<td>45.2</td>
</tr>
<tr>
<td>LSTM + One-Hot</td>
<td>0.1</td>
<td>3.5</td>
<td>30.65</td>
<td>34.65</td>
</tr>
<tr>
<td>LSTM + BERT CSL</td>
<td>0.2</td>
<td>8.3</td>
<td>22.72</td>
<td>27.4</td>
</tr>
<tr>
<td>LSTM + Google BERT</td>
<td>0.45</td>
<td>15.46</td>
<td>31.9</td>
<td>36.35</td>
</tr>
</tbody>
</table>

Juven’s CSL: Quantitative Analysis

Table 1: Accuracy for word prediction using Xception CNN classifiers. The precision and recall for both single and multi-task experiments are listed in the table.

Juven’s CSL: Qualitative Analysis

Figure 1: Echo State Networks are an instance of the Reservoir Computing paradigm using units with continuous states

Figure 2: Evaluations of different imagined scenes: Valid and Exact Errors

Figure 3: Juven’s data Fine-Tuned BERT CSL

Figure 4: Juven’s data One-Hot Encoding

Figure 5: Juven’s CSL: Output activation of the LSTM + Fine-Tuned BERT CSL

Figure 6: Juven’s Data: Output activation of the ESN + BERT CSL

Discussion

- We compare the ability of ESNs and LSTMs to learn to parse sentences via imperfect supervision (cross-situational learning)
- These experiments yield the following insights: (i) ESNs generalize better than LSTMs when the vocabulary size increases (for a comparable number of trained parameters); (ii) fine-tuned BERT representation (i.e. BERT CSL) is the best representation among all models.
- In future work, we will investigate how to transfer this surprisingly good ESN generalizing performance by adding gating mechanisms to ESNs and attention mechanisms.