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[1], Frédéric Alexandre[1], Xavier Hinaut[1]

[1]INRIA 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 working details 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)

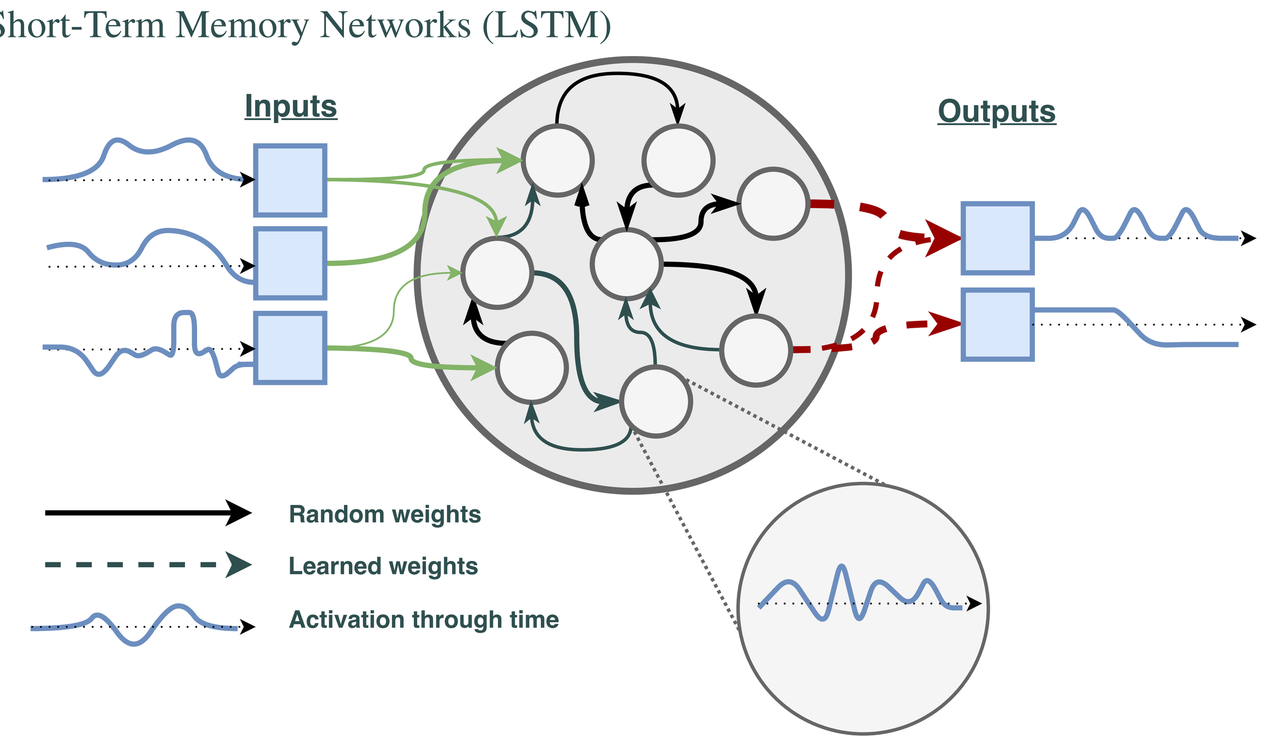

Figure 1: Echo State Networks are an instance of the Reservoir Computing paradigm using units with continuous states

Evaluation Metrics

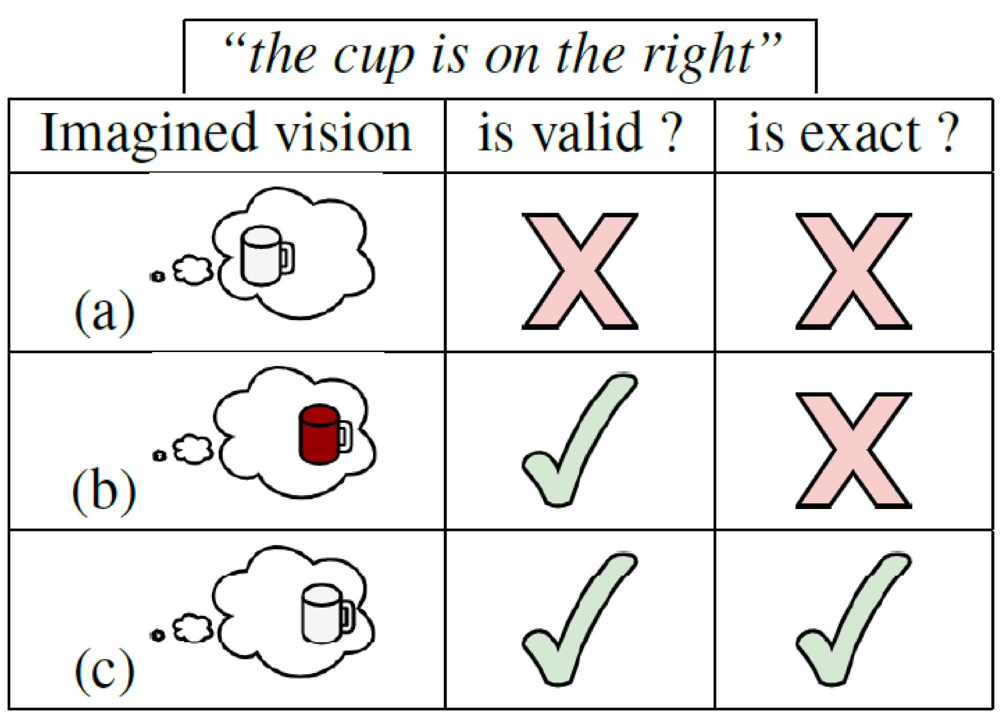

Figure 2: Evaluations of different imagined scenes: Valid and Exact Errors.

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

	Juven's CSL Data		GoLD Data	
Model	**Valid Error**	**Exact Error**	**Valid Error**	**Exact Error**
ESN FL + One-Hot	0.28	5.64	20.05	49.7
ESN FL + BERT CSL	**0**	6.28	25.22	47.4
ESN FL + Google BERT	0.2	7.72	26.8	51.48
ESN CL + One-Hot	2.32	12.1	20.16	49.5
ESN CL + BERT CSL	2.41	13.7	**18.19**	45.2
ESN CL + Google BERT	2.78	14.6	22.92	49.01
LSTM + One-Hot	0.1	3.5	30.65	34.65
LSTM + BERT CSL	0.2	**1.3**	22.72	**27.11**
LSTM + Google BERT	**0**	4.56	31.9	36.35

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

Juven's CSL: Quantitative Analysis

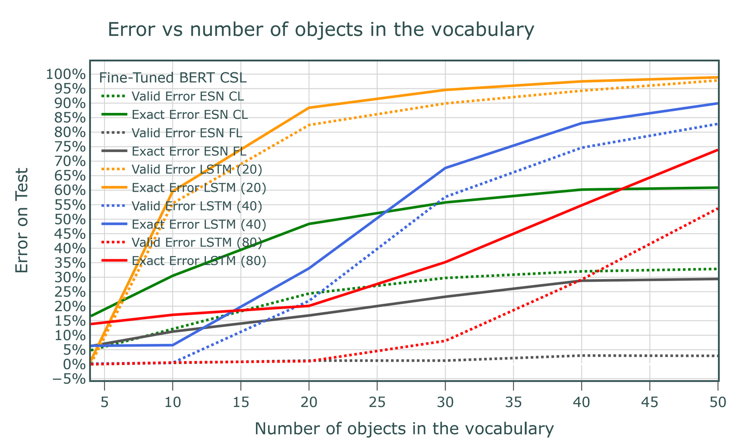

Figure 3: Juven's data Fine-Tuned BERT CSL

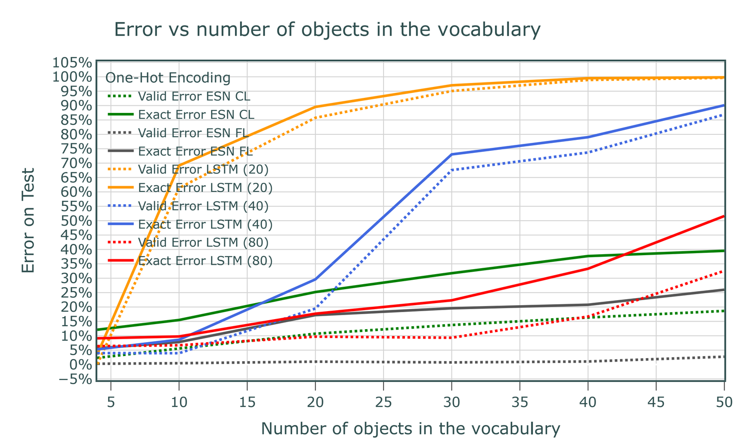

Figure 4: Juven's data: One-Hot Encoding

Juven's CSL: Qualitative Analysis

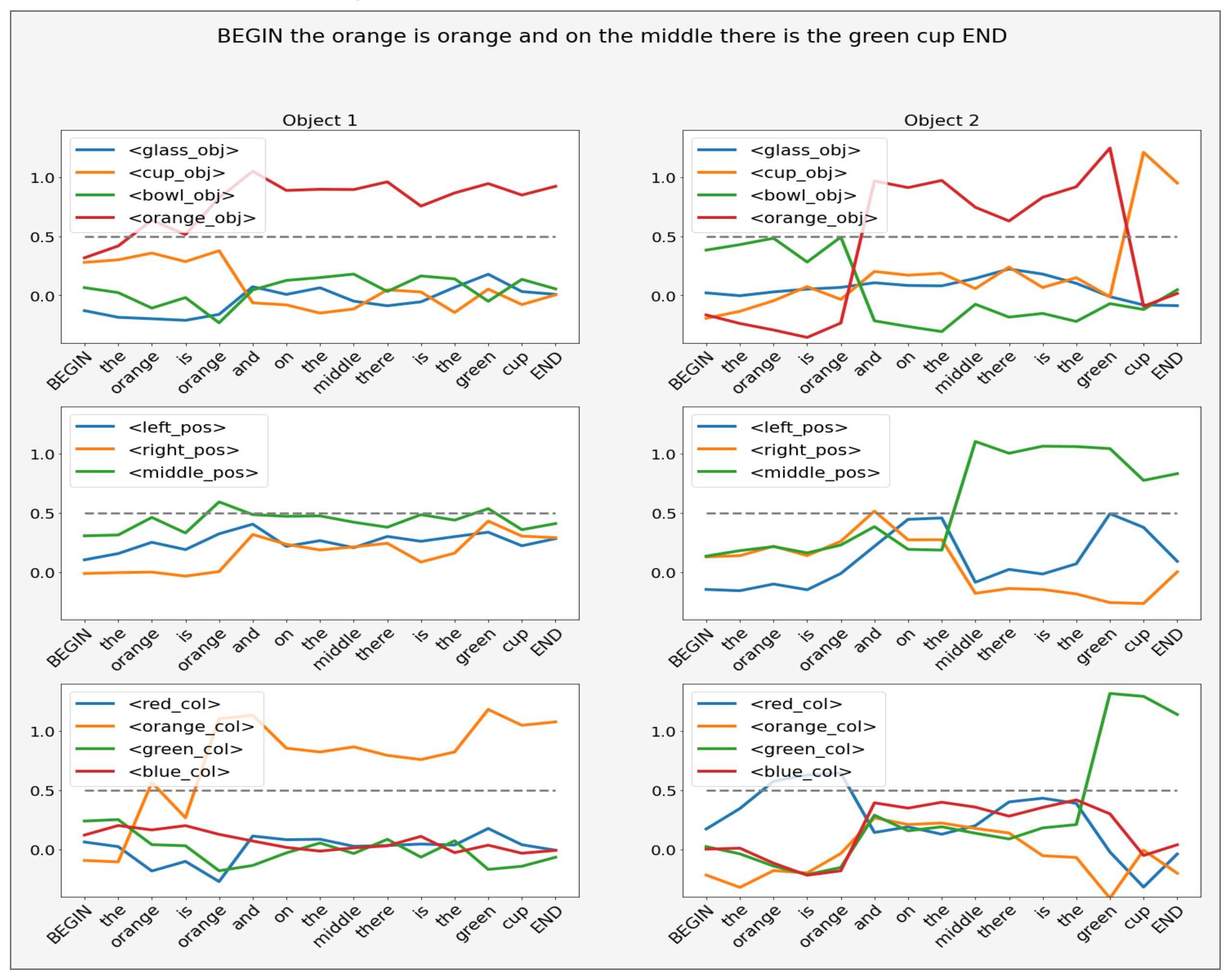

Figure 5: Juven's Data: Output activation of theLSTM + Fine-Tuned BERT CSL

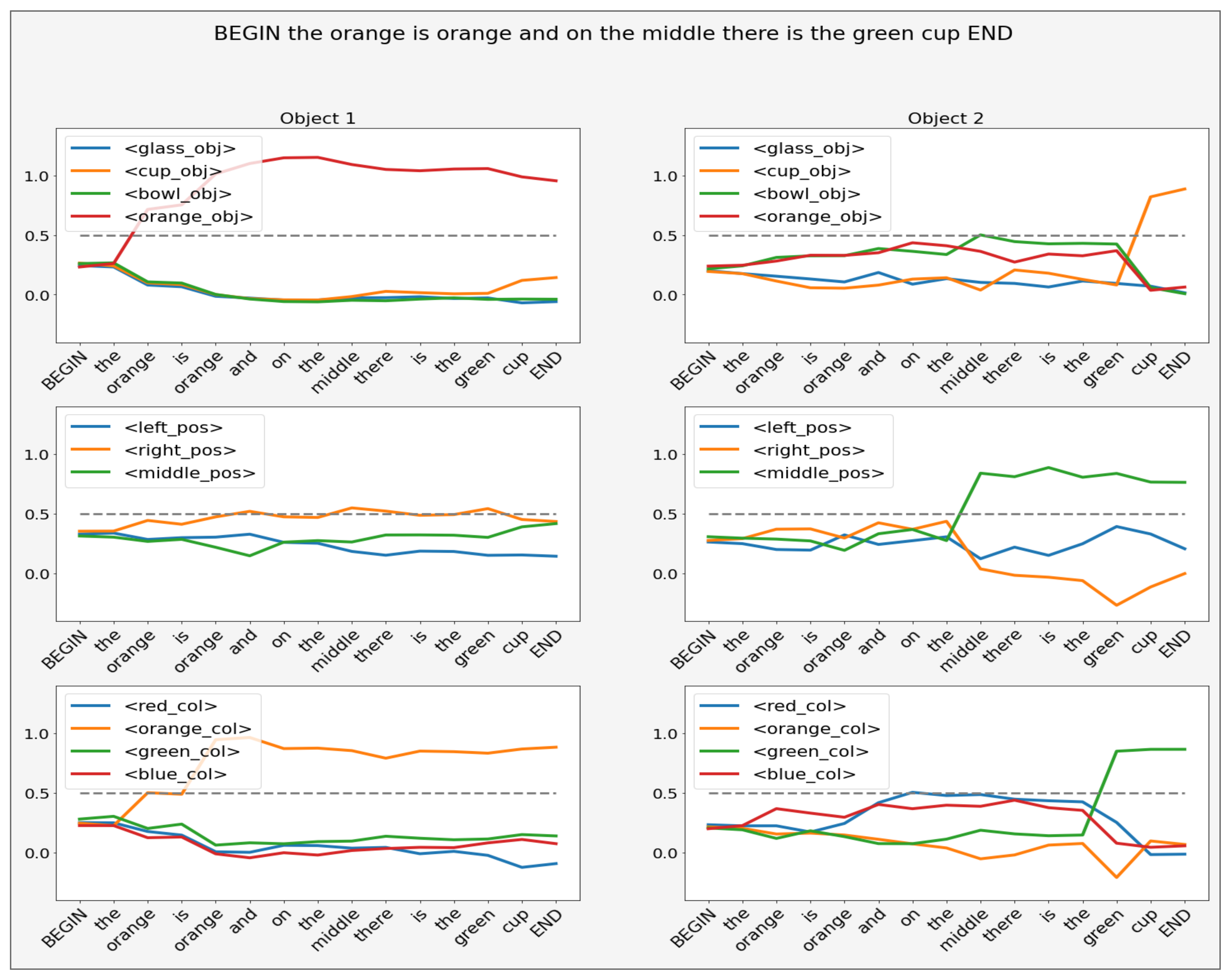

Figure 6: Juven's Data: Output activation of theESN + BERT CSL.

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.