Seminar-1

Generating and interpreting grounded language

Paper-1:

Background: According to the *physical symbol systems hypothesis* proposed by Allen Newell and Herbert Simon, human thinking is a kind of computational process (syntactic rules) done on symbols. Symbols are physical objects (values in computers, sketch on papers, sign, word) representing things in the world. For example, a symbol <APPLE> refers to an apple in the world. A symbol system is a system that consists of symbols and a set of syntactic rules that manipulates these symbols. Such can produce general intelligent action. It can be seen as a computer program capable of performing different computation processes on symbols.

The problem: The symbol system manipulates the symbols based on their shapes, not their meaning. For example, calculators do arithmetic manipulations on numbers based on their shapes, not their meaning. They do not mean anything to the calculator; they only make sense in our (human) head.

As stated in Harnad (1990), *"How can the semantic interpretation of a formal symbol system be made intrinsic to the system."* In other words, as far as we understood from the paper, Consider a robot that is controlled by a symbol system. How the robot, on its own, picks out referents for the symbols in the symbol system? So, the robot can make such connections; we should not focus only on the syntax of symbols and the rules applied to them, as the syntax operates on the shape of meaningless symbols. It is like someone who does not know the Chinese want to learn it from a Chinese/Chinese dictionary [Harnad (1990)].

Meaning Grounding: In that sense, the grounding problem goes beyond the Turing test that focuses on successful symbols manipulation. The robot must be able to detect the sensors' input features to map them to objects or world of affairs. Harnad (1990) propose the following two capacities:

1. Iconic representations or discriminate sensed objects (sensory projections of objects): to judge if two objects are the same, for example, horses in different postures.
2. Object Categorization (invariant features of objects): to identify objects, for example, identify horses from donkeys and giraffes, etc.

Neural networks are proposed to connect sensory data and symbols. Additional categories could be learned using symbolic representation: the example of zebra=horse+strips. The capacity required to interact with the environment (interface with motor) was not discussed in the paper.

Summary:

1. Symbols' shapes are arbitrary and have no connection with things they refer to
2. Symbols are manipulated synaticaly based on their shapes, not to based on what they refer to.
3. To make the system autonomously "aware" of the world it lives in; it must be able to link between the symbols and their referent in the world of affairs (Grounding problem).
4. This symbol grounding ◦s not only an innate capacity in the system but also should be learned.

Compare between a pure symbol system and a pure neural network.

	Symbol System	Neural Network
Advantages	● Explainable "Why I have this output." ● Require few data to learn	● Automatically learn features in data without the need to engineer them. ● Can deals with noise
Disadvantages	● Rules and knowledge has to be hand-coded ● Grounding problem ● Difficult to deal with noise	● Yet not explainable ● Expensive, need a lot of data. ● The tendency to learn shallow and non-related features in data

Paper-2 :

Problem: Generate spoken descriptions from visual scenes that are grounded to the user's perception.

> **Visual scene**: The image to be described consists of 10 non-overlapped colored rectangles that are randomly generated on a black background. An arrow points to a randomly selected rectangle called the target rectangle.
>
> **The task:** The system is required to perform *the rectangle description task*. The system generates contextualized spoken descriptions of the target rectangle grounded to the user's perception.

Methodology:

First: Learning algorithms are build to learn, generating unambiguous descriptions of the target object

1. word classes (color, area, shape, etc.)
2. the visual semantics of the phrase structure (features selection)
3. Order of words

Then, a planning algorithm integrates syntactic, semantic, and contextual constraints to generate natural and unambiguous descriptions of the target object.

After that, the model should learn how to generate complex utterances that describe the whole scene.

Learn visual semantics: cluster words into classes to learn which visual features are associated with a word. Three algorithms are used:

A. the first one is based on the assumption that word pairs that appear in the same utterance do not belong to the same classes. The algorithm was able to separate color terms, shape terms, size description, and brightness terms with some errors.
B. The second clustering algorithm is based on the univariate Gaussian distribution of word classes where each word is associated with a class.
C. Hybrid algorithm of the first and the second algorithms, to take into consideration both the word co-occurrences and semantic associations.

Features selection: For each word, calculate the association (KL distance) between a word (conditional model) and each individual visual feature considering the multivariate distribution**.** Features are then selected by iteratively maximizing the calculated associations. In this stage, we can conclude grounded words and ungrounded words (which are belong to all features). Each grounded word now belongs to a class (cluster) that is associated with multiple features. For each grounded word, a multivariate Gaussian model is estimated over the features associated with the grounded word's class.

Encoding Words Order (grounded language model): The model learns to generate a sequence of word-class based on the encoded-word order constraints (the probability of class-j comes after class-j)

Generating spoken language:

1. Syntactic and semantic constraints: The model generates a sequence of words based on the encoded-word order that it learned.
    a. The number of classes that form the sequence (length of sequence) can be determined by iteratively maximizing the log probability of the sequence generated by the grounded language model.
    b. From the determined sequence of classes, from each class, the word that maximizes the probability of the target object is chosen.
2. Contextual constraints: Learning the semantic and syntactic features and mapping them to a sequence of generated words only is not enough to generate an unambiguous phrase. The contextual contains here is considered to be the best competing object in the scene.

Generating complex utterances: to describe the whole scene and parsing complex utterances from the training corpus (not truncated sequence) using acquired class-based language models described above. Then, the system generates a relative spatial clause by learning to select a landmark and describe it as we discussed before. Then the spatial clause is generated using a

predefined template in the form (TARGET_PHRASE <SPATIAL PHRASE> LANDMARK_PHRASE)

The system is then evaluated using three human listeners.

Questions:

1. **In paper-1:** If the purpose is to build an intelligent system, does the symbol grounding (the connection between symbol and referent) represents the meaning (referring to the semantic triangle)?