Learning Words for HRI
IURS Summer School, Benicassim

Haris Dindo

UNIPA, Italy

20th September, 2012
Overview

Problem: Incrementally learn (i.e. generate) and maintain a grounded language model from *continuous demonstrations* with known/unknown referent

Problems to solve:

- **Semantic clustering**: How to associate words with appropriate perceptual stimuli?
- **Grammar learning**: How to learn (very simple!) grammar rules?
Terminology

- **Grounding**
  The process of connecting external percepts to internal representations

- **Semantic category (item)**
  A range of sensory inputs which can be grouped and associated with a word/symbol (e.g. a portion of the color spectrum)

- **Semantic class**
  A set of semantic categories grounded in the same sensory channel (e.g. color, shape)

- **Lexical item**
  Association between a word and its corresponding semantic category
Learning problem

A. Features Extraction

B. Semantic Clustering

C. Word-to-Meaning Association

D. Semantic Bootstrapping
Outline of the talk

1. **Semantic clustering**
   - Known referent approach: density estimation
   - Unknown referent approach: multiple-instance learning (MIL)

2. **Improving word-to-meaning associations**
   - Pseudo-syntactic information
   - Probabilistic model of language
   - Quantitative results

3. **Dialogue**
   - Knowledge base
   - Ambiguity resolution

4. **Qualitative evaluation**
   - Experimental setup
   - Demo
Approach #1: Known referent

- We assume that the referent of a phrase is known
- The meaning of each word (i.e. semantic category) is treated as a multivariate Gaussian density
- The goal is to estimate these densities from contexts which co-occur with the word
Density estimation

Hypothetical semantic category distribution: $p(f^j|w)$
Background distribution: $p(f^j|\overline{w})$

Dindo and Zambuto, 2009; Zambuto et al., IJCLR, 2010
Density estimation

\[ \varepsilon \leq \sqrt{P_1 P_2 e^{-d_{bhat}}} \]

\[ d_{bhat} = \frac{1}{8} (\mu_2 - \mu_1) \left( \frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \ln \left( \frac{\sqrt{\Sigma_1 \Sigma_2}}{|\Sigma_1| |\Sigma_2|} \right) \]
Approach #2: Unknown referent

- **Goal**: learning a grounded language model from examples without knowing the referent
- **Problem**: Each example containing the word “red” contain at least one instance of the color red, but it will also contain instances of different colors that are an additional source of ambiguity
- In order to correctly infer the meaning of the word “red”, the agent must be able to isolate from each example the proper instance, discarding all others
Multiple-instance learning (MIL)

- MIL is a variation of supervised learning for problems with incomplete knowledge.
- In MIL the labels are only assigned to bags of instances.
- Two possible bags:
  - *positive*: if at least one instance in that bag is positive.
  - *negative*: if all the instances in it are negative.
- Grounded language learning can be seen as a variation of a MIL.
Multiple-instance learning (MIL)

- Problem: unlike the classical MIL paradigm we cannot guarantee that negative bags contain negative instances only (ambiguity)!
- Possible solution: integration of the pragmatic information (e.g. pointing directions, gaze direction, ...) to filter the possible referents and weight single instances in each bag
MIL overview

Dindo and Zambuto, IROS, 2010
Negative bag ambiguity

We define a function $\phi$ that computes an estimate of “saliency” of each instance in a bag.
Objective of MIL

- $x_n$ denote a (positive or negative) bag and $x_{ni}$ is the $i$th instance in that bag
- For each semantic class-word pair, we must estimate two probability densities:
  - $p(x|c, w, \theta^+)$ (hypothetical semantic category)
  - $p(x|c, \overline{w}, \theta^-)$ (background distribution)
- The meaning of each word (i.e. semantic category) is treated as a multivariate Gaussian density
Learning negative models

- We must carefully select the bags to be used for estimation of the negative model $p(x|c, w, \theta^-)$
- A simple hint: learn negative models from examples where the degree of salience is concentrated on few objects (less ambiguous case) → ML estimation
A more difficult problem is to estimate a parametric model $p(x|c, w, \theta^+)$ from positive bags.

The knowledge of positive instance in each bag is modeled by using a set of binary hidden variables $z$

Iteratively solve by using Expectation Maximization (EM) algorithm
Learning positive models

\[ p(x \mid z_j) = N(x_j \mid \theta^+)(1 - N(x_j \mid \theta^-)) \prod_{i \neq j} N(x_i \mid \theta^-) \]

- \( p(x \mid z_j) \) represents a measure of the intersection of the positive bags minus the union of the negative bags.
Learning positive models: modified EM algorithm

**Expectation step**

\[
p(Z|X, \theta^+) \approx \prod_n \prod_j \phi_{nj} p(x_n|z_{nj})^{z_{nj}}
\]

\[
E[z_{ni}] = \frac{\phi_{ni} p(x_n|z_{ni})}{\sum_{j=1}^{I} \phi_j p(x_n|z_{nj})} = \gamma(z_{ni})
\]

**Maximization step**

\[
\mu^+ = \frac{1}{N} \sum_n \left[ \max_i \gamma(z_{ni}) \right] x_{ni}
\]

\[
\Sigma^+ = \frac{1}{N} \sum_n \left[ \max_i \gamma(z_{ni}) \right] (x_{ni} - \mu^+)(x_{ni} - \mu^+)
\]
Semantic association

How to associate words to semantic categories?

The most diverse positive and negative model (as measured by the Bhattacharyya distance) is chosen as the correct association for a given word.
Exploiting pseudo-syntactic information

Pseudo-syntactic information

Use the *position* of the words in a sentence, along with the semantic information, to improve the associations
Probabilistic model of language

- A training example then consists of a sequence of words $w_1:T_k$ (utterance) and a set of features $F_k = f^1_k, \ldots, f^M_k$ describing the (most probable) target object ($M$ semantic class)
- Latent variable $C$ (discrete) models the relationship between word $w$ and one of the feature classes $f^m$
- We want to improve our estimate of $P(w|c)$: iteratively learn model parameters (Baum-Welch algorithm)
A modified version of Baum-Welch algorithm:

**Expectation step**

\[
\epsilon_t(i, j) = \frac{\alpha_t(i) a_{ij} b_{js} \mathcal{N}(f_k^j | \mu_{js}^+, \Sigma_{js}^+)}{\sum_i \sum_j \alpha_t(i) a_{ij} b_{js} \mathcal{N}(f_k^j | \mu_{js}^+, \Sigma_{js}^+)} \beta_t(j)
\]

\[
\lambda_t(i) = p(C_t = i | w_t = s, F_k) = \sum_j \epsilon_t(i, j)
\]

**Maximization step**

\[
a_{ij} = \frac{\sum_t \epsilon_t(i, j)}{\sum_t \lambda_t(i)}
\]

\[
b_{js} = \frac{\sum_{t, w_t = s} \lambda_t(j)}{\sum_t \lambda_t(j)}
\]
Results

(a) Word-to-meaning association: Bhat-tacharyya error
(b) Word-to-meaning association: pseudo-syntactic information
Knowledge base

We use first-order logical terms to represent the state of the world

- Each object in a scene is described and added to a Prolog knowledge base
- Possible queries: `red(X)`, `above(X,Y)`, ...
- It helps decreasing the computational burden
Ambiguity resolution

How do we conduct human-robot dialogue in presence of perceptual ambiguities?

Tell me something about the blue object!
Disambiguation tree

- *Disambiguation tree* is a well-balanced decision tree.
- Each node contains a yes/no question that maximally disambiguate a given situation.
- Target objects (referents) are stored in the leaves.
- Disambiguation strategy is a path from the root to a leaf.
Disambiguation tree: the algorithm

- Constructed each time an ambiguous query is presented (logic queries with multiple answers)
- At each decision node the algorithm chooses a possible question and splits the available objects into two groups
- The question that minimizes an entropy-based measure is selected

\[
\frac{N_0}{N_1} H(X_0) + \frac{N_1}{N_0} H(X_1) + \sum_i C_i(X^k)
\]
Disambiguation tree: an example

blue(X)
  └── trapezium(X)
    │    └── obj6
    │         └── obj4
    └── circle(X)
        └── obj1
blue(X)
  └── yellow(X)
    │    └── rect(X)
    │         └── obj2
    └── below(X, obj6)
      │    └── obj7
      │         └── obj3
      └── obj5
Experimental setup

- We have tested the system on the NAO humanoid platform
- NAO uses its own speech recognition and speech production capabilities
- We have built a set of perceptual and motor schema for basic behaviors

(c) (d)
System in action

Video #1
Video #2
This work has been partially supported by the EU-FP7 funded project **HUMANOBS** (www.humanobs.org)
Thank you for your attention!