Adding Structure and Machine Learning to Speech Recognition Systems

Alex Acero

Microsoft Research

Learning Semantic CFGs

- Problem:
  - Simple CFGs tend not to have good coverage
  - High coverage CFGs are difficult to write
- Solution
  - Learning a semantic context free grammar (CFG) from example sentences and corresponding schema annotations
  - Easier authoring *AND* lower error rates

Personalization: Adding Structure

- Example: dictating names in MS Outlook
- Build CFGs for To: and Cc: fields with names in Sent Emails Folder
  - p("Bill Gates")=0.0005
  - p("Kai-Fu Lee")=0.01
- Aging: old emails have lower weight
- Huge decrease in error rate
- Semantics S: Email recipients
- Train $p(A,W,S)$ with (W,S) pairs

Learning from user corrections

- Current desktop dictation scenario
  - User dictates
  - System makes mistake
  - User highlights error and corrects it (select from alternates list, respeaks, types)
  - System will make same mistake again, and again ☹
- System should learn from user corrections:
  - Is it a new word?
  - Is dictionary pronunciation wrong?
  - Should we update the LM?
  - Should we update the acoustic model?
  - $p(A, W, \Phi_{AM}, \Phi_{DIC}, \Phi_{LM})$

How do children learn?

- They do not get (*A*,*W*) pairs
- They use semantic structure *S*
  - Child: “What’s a dove?”
  - Parent: “A dove is a bird”
  - Child: “Doves fly”
- The use vision *V* too
  - Toddlers match acoustics *A* and vision *V*
- They train *p*(*A*, *W*, *S*, *V*) with (*A*,*V*) pairs

Improving the user interface: Multimodality

- Add image and mouse to interaction
- Benefits of multimodality
  - User sees system response more quickly
  - User knows what system expects
  - User sees what system understood
- Multimodal more user-friendly than IVR

Future directions

- Adding structure can improve learning
- Learning: maximize new joint probability

  $p(A, W, S, \Phi_{AM}, \Phi_{DIC}, \Phi_{LM}, \Phi_{SEM})$
  - *A:* Acoustic signal
  - *W*: Word String
  - *S*: Semantic info (i.e. email recipient)
- Advanced learning can improve accuracy
  - Not all variables are observed