ASR, NLU, DM

Ling575
Spoken Dialog Systems
April 12, 2017

Roadmap

- ASR
  - Basic approach
  - Recent developments
- NLU
  - Call routing
  - Slot filling:
    - Semantic grammars
    - Sequence models
- DM:
  - Finite-state and Frame-based models

Summary: ASR Architecture

- Five easy pieces: ASR Noisy Channel architecture
  1) Feature Extraction:
     39 "MFCC" features
  2) Acoustic Model:
     Gaussians for computing p(o|q)
  3) Lexicon/Pronunciation Model
     - HMM: what phones can follow each other
  4) Language Model
     - N-grams for computing $p(w_i|w_{i-1})$
  5) Decoder
     - Viterbi algorithm: dynamic programming for combining all these to get word sequence from speech!

Deep Neural Networks for ASR

- Since ~2012, yielded significant improvements
- Applied to two stages of ASR
  - Acoustic modeling for tandem/hybrid HMM:
    - DNNs replace GMMs to compute phone class probabilities
    - Provide observation probabilities for HMM
  - Language modeling:
    - Continuous models often interpolated with n-gram models

DNN Advantages for Acoustic Modeling

- Support improved acoustic features
  - GMMs use MFCCs rather than raw filterbank ones
    - MFCCs advantages are compactness and decorrelation
    - BUT lose information
    - Filterbank features are correlated, too expensive for GMM
  - DNNs:
    - Can use filterbank features directly
    - Can also effectively incorporate longer context
- Modeling:
  - GMMs more local, weak on non-linear; DNNs more flexible
  - GMMs model single component; (D)NNs can be multiple
  - DNNs can build richer representations

Why the post-2012 boost?

- Some earlier NN/MLP tandem approaches
  - Had similar modeling advantages
- However, training was problematic and expensive
- Newer approaches have:
  - Better strategies for initialization
  - Better learning methods for many layers
    - See “vanishing gradient”
  - GPU implementations support faster computation
    - Parallelism at scale

Word Error Rate

- Word Error Rate =

100 (Insertions+Substitutions + Deletions)

Total Word in Correct Transcript

Aligment example:

REF: portable **** PHONE UPSTAIRS last night so

HYP: portable FORM OF STORES last night so

Eval I S S

WER = 100 (1+2+0)/6 = 50%

NIST sctk-1.3 scoring software: Computing WER with sclite

- http://www.nist.gov/speech/tools/
- Sclite aligns a hypothesized text (HYP) (from the recognizer) with a correct or reference text (REF) (human transcribed)

```
id: (2347-b-013)

Scores: (#C #S #D #I) 9 3 1 2

REF:  was an engineer SO I   i was always with **** **** MEN UM   and they

HYP:  was an engineer ** AND i was always with THEM THEY ALL THAT and they

Eval:                 D  S                     I    I    S   S
```

Better metrics than WER?

- WER has been useful
- But should we be more concerned with meaning ("semantic error rate")?
  - Good idea, but hard to agree on
  - Has been applied in dialogue systems, where desired semantic output is more clear

Accents: An experiment

- A word by itself

- The word in context

Challenges for the Future

- Doing more with more
  - More applications:
    - From Siri, in-car navigation, call-routing
    - To full voice search, voice-based personal assistants, ubiquitous computing
  - More speech types:
    - Accented speech
    - Speech in noise
    - Overlapping speech
    - Child speech
    - Speech pathology

NLU for Dialog Systems

Natural Language Understanding

- Generally:
  - Given a string of words representing a natural language utterance, produce a meaning representation
- For well-formed natural language text (see ling571),
  - Full parsing with a probabilistic context-free grammar
    - Augmented with semantic attachments in FOPC
    - Producing a general lambda calculus representation
- What about spoken dialog systems?

NLU for SDS

- Few SDS fully exploit this approach
- Why not?
  - Examples of travel air speech input (due to A. Black)
    - Eh, I wanna go, wanna go to Boston tomorrow
    - If its not too much trouble I'd be very grateful if one might be able to aid me in arranging my travel arrangements to Boston, Logan airport, at sometime tomorrow morning, thank you.
    - Boston, tomorrow

NLU for SDS

- Analyzing speech vs text
  - Utterances:
    - ill-formed, disfluent, fragmentary, desultory, rambling
      - Vs well-formed
  - Domain:
    - Restricted, constrains interpretation
      - Vs. unrestricted
  - Interpretation:
    - Need specific pieces of data
      - Vs. full, complete representation
  - Speech recognition:
    - Error-prone, perfect full analysis difficult to obtain

NLU for Spoken Dialog

- Call routing (aka call classification):
  - (Chu-Carroll & Carpenter, 1998, Al-Shawi 2003)
  - Shallow form of NLU
  - Goal:
    - Given a spoken utterance, assign to class c, in finite set C
  - Banking Example:
    - Open prompt: **"How may I direct your call?''**
    - Responses: may I have consumer lending?,
      - I'd like my checking account balance, or
      - "ah I'm calling 'cuz ah a friend gave me this number and ah she told me ah with this number I can buy some cars or whatever but she didn't know how to explain it to me so I just called you you know to get that information."

Call Routing

- General approach:
  - Build classification model based on labeled training data, e.g. manually routed calls
    - Apply classifier to label new data
- Vector-based call routing:
  - Model: Vector of word unigram, bigrams, trigrams
    - Filtering: by frequency
      - Exclude high frequency stopwords, low frequency rare words
    - Weighting: term frequency * inverse document frequency
    - (Dimensionality reduction by singular value decomposition)
  - Compute cosine similarity for new call & training examples

Meaning Representations for Spoken Dialog

- Typical model: Frame-slot semantics
  - Majority of spoken dialog systems
    - Almost all deployed spoken dialog systems

- Frame:
  - Domain-dependent information structure
  - Set of attribute-value pairs
  - Information relevant to answering questions in domain

Natural Language Understanding

- Most systems use frame-slot semantics
  Show me morning flights from Boston to SFO on Tuesday

  - SHOW:
  - FLIGHTS:
    - ORIGIN:
      - CITY: Boston
      - DATE:
        - DAY-OF-WEEK: Tuesday
      - TIME:
        - PART-OF-DAY: Morning
    - DEST:
      - CITY: San Francisco

Another NLU Example

- Sagae et 2009
- Utterance (speech): we are prepared to give you guys generators for electricity downtown
- ASR (NLU input): we up apparently give you guys generators for a letter city don town
- Frame (NLU output):
  - **<s>.mood declarative**
  - **<s>.sem.agent kirk**
  - **<s>.sem.event deliver**
  - **<s>.sem.modal.possibility can**
  - **<s>.sem.speechact.type offer**
  - **<s>.sem.theme power-generator**
  - **<s>.sem.type event**

Question

- Given an ASR output string, how can we tractably and robustly derive a meaning representation?
- Many approaches:
  - Shallow transformation:
    - Terminal substitution
  - Integrated parsing and semantic analysis
    - E.g. semantic grammars
  - Classification or sequence labeling approaches
    - HMM, MaxEnt, CRF, sequence NNs

Grammars

- Formal specification of strings in a language
- A 4-tuple:
  - A set of terminal symbols: Σ
  - A set of non-terminal symbols: N
  - A set of productions P: of the form A -> α
  - A designated start symbol S
- In regular grammars:
  - A is a non-terminal and α is of the form $\{N\}\Sigma^*$
- In context-free grammars:
  - A is a non-terminal and α in $(\Sigma \cup N)^*$

Simple Air Travel Grammar

- LIST -> show me | I want | can I see|…
- DEPARTTIME -> (after|around|before) HOUR| morning | afternoon | evening
- HOUR -> one|two|three…|twelve (am|pm)
- FLIGHTS -> (a) flight|flights
- ORIGIN -> from CITY
- DESTINATION -> to CITY
- CITY -> Boston | San Francisco | Denver | Washington

Shallow Semantics

- Terminal substitution
  - Employed by some speech toolkits, e.g. CSLU
- Rules convert terminals in grammar to semantics
  - LIST -> show me | I want | can I see|…
    - e.g. show -> LIST
    - see -> LIST
    - | -> ε
    - can -> ε
    - * Boston -> Boston
- Simple, but…
  - VERY limited, assumes direct correspondence

Semantic Grammars

- Domain-specific semantic analysis
- Syntactic structure:
  - Context-free grammars (CFGs) (typically)
    - Can be parsed by standard CFG parsing algorithms
      - e.g. Earley parsers or CKY
- Semantic structure:
  - Some designated non-terminals correspond to slots
    - Associate terminal values to corresponding slot
- Frames can be nested
- Widely used: Phoenix NLU (CU, CMU), vxml grammars

Show me morning flights from Boston to SFO on Tuesday

- LIST -> show me | I want | can I see|…
- DEPARTTIME -> (after| around|before) HOUR| morning | afternoon | evening
- HOUR -> one|two|three…| twelve (am|pm)
- FLIGHTS -> (a) flight|flights
- ORIGIN -> from CITY
- DESTINATION -> to CITY
- CITY -> Boston | San Francisco | Denver | Washington

- SHOW:
- FLIGHTS:
  - ORIGIN:
    - CITY: Boston
    - DATE:
      - DAY-OF-WEEK: Tuesday
    - TIME:
      - PART-OF-DAY: Morning
  - DEST:
    - CITY: San Francisco

Semantic Grammars: Issues

- Issues:
  - Generally manually constructed
    - Can be expensive, hard to update/maintain
  - Managing ambiguity:
    - Can associate probabilities with parse & analysis
    - Build rules manually, then train probabilities w/data
  - Domain- and application-specific
    - Hard to port

Learning Probabilistic Slot Filling

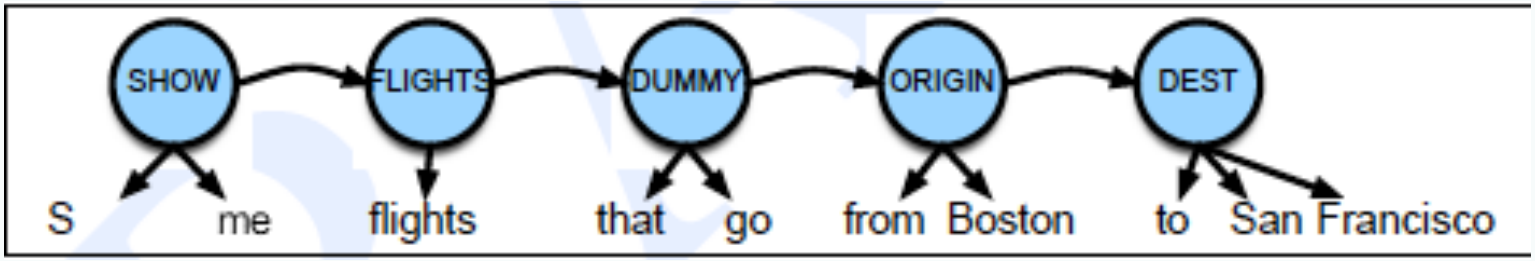
- Goal: Use machine learning to map from recognizer strings to semantic slots and fillers
- Motivation:
  - Improve robustness – fail-soft
  - Improve ambiguity handling – probabilities
  - Improve adaptation – train for new domains, apps
- Many alternative classifier models
  - HMM-based, MaxEnt-based, CRF-based
  - DNN sequence models: RNN/LSTM

HMM-Based Slot Filling

- Find best concept sequence C given words W
- C^* = argmax P(C|W)
- = argmax P(W|C)P(C)/P(W)
- = argmax P(W|C)P(C)
- Assume limited M-concept history, N-gram words
  - = $$\prod_{i=2}^{N} P(w_i \mid w_{i-1} \ldots w_{i-N+1}, c_i) \prod_{i=2}^{N} P(c_i \mid c_{i-1} \ldots c_{i-M+1})$$

Probabilistic Slot Filling

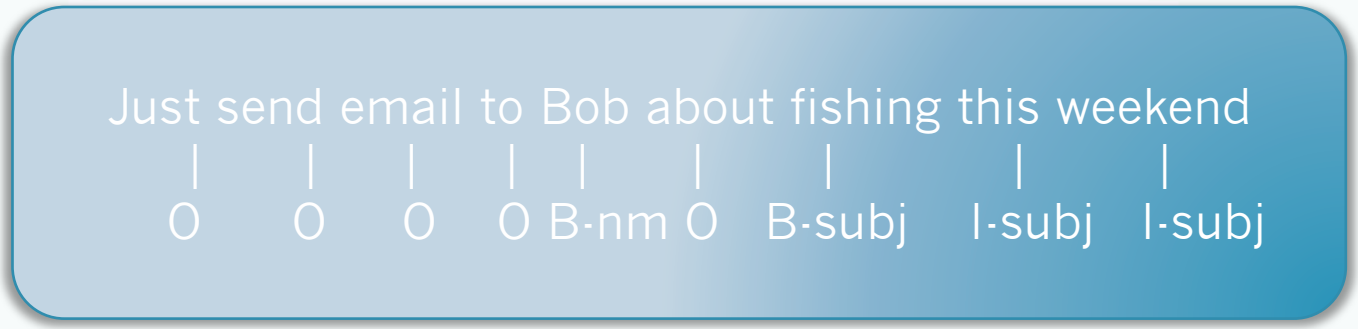
- Example HMM



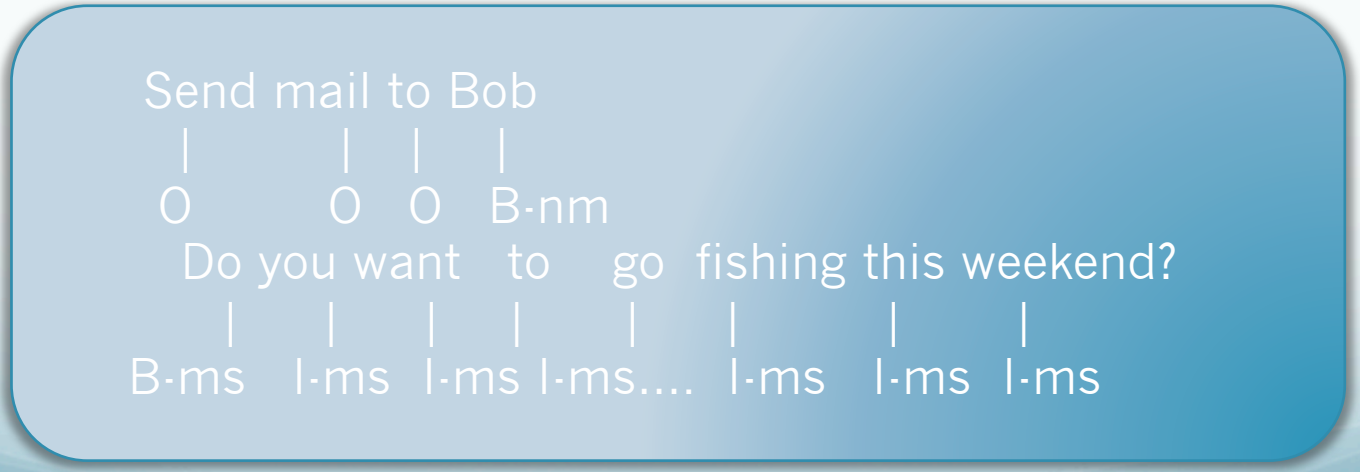
Joint Sequence Models

- NLU tasks in a multi-domain setting require:
  - Domain: e.g. communication, reminder, music, etc
  - General “intent”: send-mail, read-mail, call-friend
  - Slot-filling: recipient=bob; message = “blah blah blah”
- Classically done in series:
  - Domain → intent → slots
- Issues?
  - Error propagation
  - Ignore mutual constrains

Sequence Labeling Task



- Domain: comm; Intent: send_email



Approaches

- Almost any sequence model could be applied
  - And has been
  - HMM, HMM-SVMs, CRFs
  - Incorporate evidence from current word observation as well as previous state
    - Alternate models of dependency
- Recent work neural sequence models
  - RNNs (recurrent NNs)
  - LSTMs (Long Short-term memory models)

Joint Context Models

- Incorporate longer term history
  - E.g. multiple speaker turns
- Output joint prediction of intent, slot values
- Models:
  - Current and prior turns as embeddings
  - History encoded as attention-weighted sum
  - Sequence classification as RNN with gated recurrent units (GRUs)
    - Interpolates between output at prior time step & current

Specialized Topics

- Dialog for different pops
  - Universal access
- Context modeling:
  - Discourse/Anaphora
- Miscommunication/Repair
- Grounding
- Prosody in dialog
- Incremental processing
- Turn-taking/backchannels
- Entrainment
- Emotion/affect/sentiment
- Knowledge acquisition
- Domain adaptation
- Ethics in SDS
- Applications:
  - Language teaching
  - Medical/Therapeutic,
  - Voice search/QA, etc
- Interactional dialog/chatbots
- Persona/personality
- Systems:
  - Generation & TTS for dialog
  - NLU/slot-filling/intent
  - ASR (& phonology/phonetics)
  - Evaluation
  - Shared tasks (DSTC)
- Multi-party systems
- Multi-modal systems
- Multi-linguality

Specialized topics: To-do

- Reply to GoPost thread with preferred topics
- Presenter:
  - Will read/synthesize/discuss ~ papers (10-15 min)
  - Lead discussion (10 min)
- Participants:
  - Read a paper from each topic
  - Select one to write a “critical response” 1-page paper
  - Submit at least one question to GoPost