Computational Methods in Linguistics

Bender and Wassink 2012
University of Washington
Week 2: 4/2-4/4

Practicum Groups

- Meet this week outside of class to brainstorm research questions
- Brief reports on Wednesday of the results

Some personal history

- LSA 2009 Symposium: Computational methods in support of Linguistic analysis (January)
- Cyberling 2009 workshop (July)
- Special issue of LILT (2010)
- Cyberling blog...

Bender & Langendoen 2010: pie in the sky

- lots and lots and lots of accessible data
- facilities for analyzing data & enriching it with the results
- support for collaboration across time, space, (sub)disciplines, and theories

Bender & Langendoen 2010: To do now

- Share data
- Teach:
  - What resources exist, what standards/ best practices exist, corpus manipulation tools, db querying, high-level programming, general computational skills
- Effect culture change

Reading Questions

- How to pool together data for metalinguistic commentary, folk linguistics, perceptual dialectology
- What would an annotation scheme for metalinguistic commentary look like?
- How to distribute data if this wasn't foreseen in original IRB application; considering different aspects of the data separately

Reading Questions

- What does that meta-language look like?
- What about SLA CALL system users' data as a data source?
- How could ELTK be extended to detect probable sites of variation?
- What are DILI and DIGPI? Actual resources?
- How can we best take advantage of human-machine collaboration?

Automatic annotation

- Various NLP tools exist that can (with a certain amount of noise) add annotations to text that may be useful in linguistic research

What do we wish to accomplish with our database?

- What are you working on now?
- How can we use linguistic databases to further those goals?

NLP tools

- POS taggers
- Constituency parsers
- Dependency parsers
- ASR
- Forced alignment

How are these built?

- Knowledge-based systems involve hand-coded rules that linguists have put together over many person-years of effort.
- Tend to be precise, but not always robust

http://erg.delph-in.net

How are these built?

- Stochastic or statistical systems are built via *machine learning*
- Machine learners are computer programs that extract patterns from sets of *training data* and then use those patterns to predict labels for additional data

Training data

- For *supervised* machine learning the training data is labeled.
  - Patterns are learned between the labels and other properties of the data
  - Training data comes from annotation projects, i.e., human labor
- Ex: The Stanford parser (Klein and Manning 2003), trained on the Penn Treebank (Marcus et al 1993)

http://nlp.stanford.edu:8080/parser/

Creating training data involves annotating running text

- Annotation schemes need to be developed and refined in light of what's actually in the data
- For many types of annotation (e.g., POS tagging) the annotation scheme has to have complete coverage

POS tags

- How many different parts of speech are there in English?
- Step 1: brainstorm without looking at data
- Step 2: look at small sample of text
- Step 3: look at PTB tag set
- Why do we as consumers need to understand the whole PTB tag set?

Step 1: Brainstorm POS tags

Noun
Verb
Adjective
Adverb
Article
Preposition
Auxiliary
Negation
Conjunction
Complementizer
wh words
Quantifiers

Valence: trans, intrans
Form: gerund, present, past

Step 2: test them

The Huskies Are Once Again Top Dog

After a long, long stretch at the top of the Graduate Tournament leaderboards, Stanford University has been deposed! The University of Washington has reclaimed its former status as top school, showing an astounding $4005 in donations. With perhaps as little as three days remaining in Fund Drive, can anyone overtake U. Washington? Will another school school snag the Golden Pig Award for themselves?

http://linguistlist.org/issues/23/23-1630.html

Step 3: see what others have done

- PTB tag set
- Extended tags

Speech tools

- Automatic speech recognition, i.e., automatic orthographic transcription
  - Input: recording output: text
- Forced alignment:
  - Input: recording, transcription; output: alignment between the two

More on ASR (from Levow's slides)

- ASR inputs:
  - Speech waveform
  - Acoustic models of speech sounds
  - Pronunciation dictionary using those sounds
  - Model of word sequences in language

More on ASR

- $\hat{i} = \text{argmax}(p(i|o))$
  - given observed acoustic signal, what is the most likely input/intended string?
- $\hat{i} = \text{argmax}(p(o|i)p(i))$ [Bayes' rule]
  - string which maximizes own likelihood time likelihood of observed acoustics
- language model * acoustic model

ASR models

- language model trained on variety of corpora, "in domain" will work better than "out of domain"
- acoustic model trained with particular speakers; acoustic models trained with closer varieties will work better
- single speaker is easier than multi-speaker
- recording conditions matter

More on forced alignment (Levow)

- Forced alignment inputs:
  - Speech waveform
  - Acoustic models of speech sounds
    - Pronunciation dictionary using those sounds
    - Exact word-level transcription of speech in wave-form
- Searches for best alignment of words (and phones) to time segments in wave form

Why forced alignment (Levow)

- Fine-grained manual word alignment time-consuming
  - Fine-grained manual phone alignment is worse
    - And requires more expertise
- Word transcription is faster, easier
  - Transcriptions often available with recorded audio (e.g., closed captioning)
- Forced alignment
  - Used in ASR research to speed data acquisition
  - Used in other speech research to
    - Speed transcription and alignment (~10 x)
    - Leverage transcription effort from less-trained speakers

Extending the dictionary (Levow)

- Issues:
  - Word not in dictionary at all
  - Word in dictionary but not with desired pronunciation
- Solution: Add words to dictionary
- CMUdict:
  - Orthography: All caps
  - Pronunciation: stress marked ARPABET
- On patas, you'd want to make your own copy of the p2fa directory to add to the dictionary

Questions to ask about automatic annotators

- Where does the information come from?
- How does it relate to what linguists say about the structures in question?
- How is this information useful to you?
- What will noise in the annotations look like?
- How will noise affect your use of the annotations?