Learning linguistic structure from linguistic data

Sharon Goldwater
How can a computational system (whether human or machine) learn **linguistic structure** from **linguistic data**?
Linguistic structure

V: [see]
Det: [the]
N: [doggies]

meaning

- syntax
- words
- morphemes
- phones
- acoustics
Linguistic data

• Mostly, phones or words (spoken or written)
• Recently, acoustics and social contexts
• i.e., unsupervised
  – Like kids
  – NLP for new languages
  – Challenging/interesting
  – Useful ML models
Word segmentation

where’s the doggie
Word segmentation - idealized

werz  Marketable
(Where’s the doggie)

werz门窗
(Wheresthedoggie)
Approach

• Input:

lookatthedoggie
wheresthedoggie
yeahlookatthat
hescomingtogetyou
whatabigdoggie
didhelookatyou
Approach

• Input:

  look at the doggie
  where is the doggie
  yeah look at that
  he is coming to get you
  what a big doggie
  did he look at you
Approach

• Input:

look at the doggie
where is the doggie
yeah look at that
he is coming to get you
what a big doggie
did he look at you
Approach

• Input:

lookatthedadog
whereisthedog
yeahlookatat
hescomingtogetyou
whatabigdog

didhe
lookatyou

• Problems:
  – Common word sequences are coherent
  – How many vocabulary items?
Approach

• Input:

lookatthedoggie
wheresthedoggie
yeahlookatthat
hescomingtogetyou
whataborbigdoggie
didhe\textcolor{blue}{lookaty}ou

• Problems:

– Common word sequences are coherent
– How many vocabulary items?

• Solutions:

– Use a nonparametric Bayesian model
– and learn bigram probabilities: $P(w_i | w_{i-1})$
Results

• Compared to previous work,
  – More accurate segmentation
  – Closer match to human data
• Model and extensions later used in
  – information extraction
  – machine translation
  – native language identification
  – syntactic parsing
Now: acoustic word segmentation
Acoustic variability

- Variability within speaker
- Variability across speakers

Look at the doggie
Where’s the doggie
Yeah, look at that
Representing speech

• Standard method:

[24.5, 0.23, 3.2, ... , 0.15]
Learning better representations

- Classifies test items better with less training data
- Project: further experiments, other domains
Cognitive science aspects

• What are infant’s word representations like?
  – various proposals but often vague
  – **Project:** model the development of infant lexicon and compare proposals to human data
  – **Project:** investigate our new representations too
Bootstrapping annotated data

• At least 6500 languages in the world.
  – Many near extinction, others trying to revive; also many widely-spoken but unwritten languages.
  – Transcribing and annotating data helps linguists and speakers.

• Can we use our methods to aid annotation?
  – Active learning
  – Projects: visualization, active learning

http://www.linguisticsociety.org/content/how-many-languages-are-there-world
Conclusion

• Lots of interesting work in this space, for lots of different backgrounds!