LANGUAGE MODELS

Why should agents do NLP?

- Knowledge acquisition from spoken and written language artifacts (e.g. on the web)
  - *Natural* language is messy!

- Communicate with human

Outline

- Language Models
  - Predict the probability distribution of language expressions

Language Models

- Formal languages (e.g. Python, Logic)
  - Grammar (generative)
  - Semantics

- Natural languages (e.g. English)
  - Grammaticality is less clear
    - ** To be not invited is sad*
  - Ambiguity at many levels (syntax, semantics, …)
    - *I saw the man with the telescope*
    - *He saw her duck*
  - Suggests modeling via probability distributions
    - What is the probability that a random sentence would be a string of words?
    - What is the probability distribution over possible meanings for a sentence?

N-Gram Models

- N-Gram
  - a sequence (of some unit – characters, words, etc.) of length n
  - Unigram, Bigram and Trigrams for n= 1, 2, and 3

- N-Gram Model
  - probability distribution of n-unit sequences
  - Markov chain of order n -1
    - the probability of a unit depends only on some of the immediately preceding units

N-gram character models

- $P(c_{1:n})$ is the probability of a sequence of N *characters* c_1 through c_N
  - Typically corpus-based (uses a body of text)
  - P("the") = .03
  - P("zgq") = .000000000002

- Application: language identification
  - Corpus: P(Text | Language) (trigrams)
  - Language Identification – use Bayes Rule!

- Application: named–entity recognition
  - "ex" -> drug name
  - Can handle unseen words!

Smoothing

- What do we do about zero (or low) counts in a training corpus?
  - Sequences with count zero are assigned a small non- zero probabililty (support generalization)
  - Need to adjust other counts downward, so probability still sums to 1
- Add one smoothing (1/(n+2))
- Backoff (e.g. if no trigram, use bigram)
- Many others in NLP course
- Just like ML, is it better to improve smoothing methods, or to get more data???

Evaluation

- Just like ML, cross-validation with train/validate/test data
- Just like ML, many metrics
  - extrinsic – e.g. language identification
  - instrinsic - perplexity

N-gram *word* models

- Much larger "vocabulary" of units
- Since units are open, out of vocabulary becomes a problem
- "Word" needs to be defined precisely
- Common in speech recognition