Probabilistic Modeling of Language: An Introduction and Apologia

Dan Jurafsky

Department of Linguistics, Department of Computer Science,
Institute of Cognitive Science &
Center for Spoken Language Research

University of Colorado, Boulder

Suggestive Facts

- Language and speech input is noisy, ambiguous, and unsegmented
- Other cognitive sciences use probabilistic/statistical models to deal with such problems:
  - human visual processing (Rao et al. 2001; Weiss & Fleet 2001)
  - categorization (Tenenbaum 2000; Tenenbaum and Griffiths 2001b; Tenenbaum and Griffiths 2001a)
  - human understanding of causation (Rehder 1999; Glymour and Cheng 1998)
- Why?
  - Probability theory is good normative model for solving problems of decision-making under uncertainty
  - Humans certainly get enough input data to do statistical modeling (Baayen estimate of 200 million words)

But maybe language is irrational?

- Perhaps probability/statistics is a good normative model, but bad descriptive one of language?
- Perhaps human language is simply a non-optimal, non-rational process?
- Au contraire: We think human language is rational (in the sense of John Anderson 1990) and based on statistical mechanisms
- Different people on this panel will argue for different kinds of statistical mechanisms
- What I think they have in common:
  - Ability to model gradient effects
  - A learning theory that learns from the statistics of the input data
  - Ways of accounting for linguistic structure effects

Objection Number 291:

How could frequencies be a model of anything? Everyone knows that words or phones or constraints behave differently in different contexts, and frequencies aren't different in different contexts!

Conditional Probabilities

- Probabilistic and statistical models aren't just based on raw frequencies
- For example, probabilistic modeling makes use of two kinds of probabilities:
  - Unconditional (prior) probabilities: P(word X), or P(structure X)
  - Conditional (posterior) probabilities: P(word X $|$ context Y), or P(structure X $|$ other structure Y)
- Given that context Y occurred, what is the probability that what we just heard was word X (versus word Z)?
- Given that we've got structure X in our input, what's the probability that we will also have structure Y?

Objection Number 542:

Probability isn't really what's going on. There are some other deeper factors that really explain things and the *true* model is based on those factors, not probabilities.

Probability is not really about numbers; it is about the structure of reasoning

Glenn Shafer, cited in Pearl (1988)

- Yes, sometimes scientists don't have complete model.
- But we're using probability because sometimes the actual language user doesn't have a complete model!
- Probabilistic or statistical models explain how *people* perform linguistic acts with incomplete information.
- Probability theory says:
  - If you have ever have to choose between things
  - Compute the probability of each choice given everything you know
  - And pick the most likely one

When do we choose between multiple things?

- Comprehension:
  - Segmenting speech input
  - Lexical ambiguity
  - Syntactic ambiguity
  - Semantic ambiguity
  - Pragmatic ambiguity
- Production: choice of words (or syntactic structure or phonological form or etc)
- Learning: choosing between:
  - Constraint rankings
  - Settings of parameters
  - Different grammars
  - Possible lexical entries for new words

Objection Number 193:

But computing probabilities of really complex things is impossible because they never occur often enough to count!!!

Breaking down probabilities

Independence: Use linguistic intuitions to help come up with independence assumptions.

- Independence assumption: compute the probability of a more complex event by multiplying the probabilities of constituent events.
- *Syntax* (Bod 2003): can't compute probabilities of whole complex parse tree just by counting (too rare). So assume that the pieces are independent, and multiply probability of tree fragments.
- *Phonology* (Pierrehumbert 2003): can't compute probabilities of triphone events (too rare). So assume pieces are independent, and multiple probability of diphones.

Bayes Rule : sometimes it's easier to compute something else (generative model):

$$P(\text{structure}|\text{input}) = \frac{\overbrace{P(\text{input}|\text{structure})}^{\text{likelihood}}\overbrace{P(\text{structure})}^{\text{prior}}}{P(\text{input})}$$

Objection Number 144:

Sure, sure, maybe for engineering tasks you need all that probability stuff, because computers are stupid. But I bet there's no evidence that *people* actually compute probabilities of linguistic knowledge?

The Lexicon: Lexical Probabilities in Comprehension

V Howes and Solomon (1951): tachistoscopic presentation of iteratively longer duration; HF words recognized with less presentation.

V Forster and Chambers (1973): HF word named more rapidly.

V Rubenstein et al (1970): Lex. dec. faster to HF words.

V Howes (1957): Words masked by additive noise. High-frequency words identified better.

A Savin (1963): recognition errors biased toward words higher in frequency than presented words.

A Grosjean (1980) gating, HF words recognized earlier

A Replicated crosslinguistically, for example Tyler (1984) in Dutch.

Many other methods, including fixation, gaze duration, recall.

The Lexicon: Lexical Probabilities in Production

- Fidelholz (1975), Hopper 1976: HF words (*forgive* more likely to have reduced vowels than LF words (*forfend*)
- Bybee (2000): word-final /t/ and /d/ (in corpus of spoken English) more likely to be deleted for HF words (54.5%) than LF words (34.3%).
- Gregory *et al.* (1999) and Jurafsky *et al.* (2001): After controlling for segmental context, ROS, # phones, etc, HF words are 18% shorter than LF words.
- Oldfield and Wingfield (1965): Picture naming; pictures with HF names named faster than LF names (Wingfield (1968) showed effect is caused by word frequency, not object frequency)

Morphology: Comprehension

- Jurafsky (1996): LF syncats cause garden path:

  (1) The complex houses married and single students and their families. (*complex/A* $>$ *complex/N* and *house/N* $>$ *house/V*)

- Burgess and Hollbach (1988), Trueswell (1996): Participle/Preterite ambiguity

  Selected: 89% participle, 11% simple past
  Searched: 22% participle, 78% simple past

- In MV/RR ambiguity, preterite-bias verbs cause longer reading time for RR interpretation.

Syntax: Comprehension

- Subcategorization probabilities of verbs play role in reading time (Clifton, Jr. *et al.* 1984; Trueswell *et al.* 1993; Jennings *et al.* 1997; MacDonald 1994)
- Reading time for ambiguous word sensitive to probabilities assigned by previous syntactic structure (Juliano and Tanenhaus 1993)
- Syntactic probabilities play role in reading time for garden-path sentences (Jurafsky (1996), Corley and Crocker (1996, 2000), Narayanan and Jurafsky (1998, 2001), Hale (2001))

Syntax: dependency (word-to-word) probabilities in Comprehension and Production

- MacDonald (1993) HF noun-noun pairs (“miracle cures”) biased reading of homophone “cures” to noun; LF pairs didn’t
- McDonald, Shillock and Brew (2001): high probability words have shorter eye fixation in reading.
- Bush (1999), Bybee (1995/2000): Palatalization of coronals is more likely between higher probable word sequences (*did you*, *would you*) than less probable (*at you*, *but you*)
- Gregory *et al.* (1999), Jurafsky *et al.* (2001), Bell *et al.* (2003): High probability words more reduced (shorter, reduced vowels, deleted codas).

Phonology: Phonotactic Probability Effects

- Coda and onset probabilities explain cluster distribution in English (Pierrehumbert 1994)
- Eight month old infants are sensitive to phonological transition probabilities in artificial language speech streams (Saffran *et al.* 1996)
- Phonological parse log-probability of nonce words correlates with acceptability (Coleman and Pierrehumbert 1997)
- Phonotactic frequencies explain well-formedness judgments on nonce word and performance on blending task (Treiman *et al.* 2000)
- Cluster frequency explains transcription error rates of NO cluster (Hay, Pierrehumbert, and Beckman, in press)

Summary: Converging Evidence for Probabilistic Models in Comprehension

- Lexeme frequencies (Tyler 1984; Salasoo and Pisoni 1985; inter alia
- Lemma frequencies (Hogaboam and Perfetti 1975; Ahrens 1998;
- Idiom frequencies (d'Arcais 1993)
- Phonological probabilities Pierrehumbert 1994, Hay, Pierrehumbert and Beckman (in press), Pitt and McQueen (1998)
- Dependency (word-word) probabilities MacDonald (1993), Bod (2001), McDonald, Shillock and Brew (2001)
- Lexical category frequencies (Burgess and Hollbach 1988; MacDonald 1993, Trueswell et al. 1996; Jurafsky 1996)
- Constructional frequencies (Croft 1995; Mitchell *et al.* 1995; Jurafsky 1996, DeSmedt (2001))
- Subcategorization probabilities Ford, Bresnan, Kaplan (1982);Clifton, Frazier, Connine (1984), Trueswell *et al.* (1993)

Objection Number 145:

OK, now I understand why people would want to compute probabilities to help with disambiguation in comprehension. But why probabilities in production?

Why Probability in Production: Information Theory and Information Structure

- Why should high-probability words be lenited in production?
  - Speaker is modeling what the hearer can “figure out” (Jespersen 1923)?
  - Lexical priming for speaker? (Bard et al. 1999, Horton and Keysar 1996)
- It turns out speakers are in fact sensitive to hearer knowledge (Gregory 2001; Gregory, Healy, Jurafsky 2003)
- High probability words have low entropy, hearers need less information to interpret.
- Information value of utterance is key to choice of syntactic structure, words, phonological form.
- Bottom line: speakers are computing probability p because is useful in computing information value of utterance ($\sum p \log p$)

Objection Number 806:

OK, OK, maybe probabilities matter in language comprehension and production, but surely statistical models of learning can’t account for language learning. Language structure has to be innate. What about the famous Gold’s theorem that says that context-free grammars are unlearnable without negative evidence?

Probabilities and Learning

- Learning is really where statistical models shine.
- Today's talks (Bod, Boersma, Elman, Pierrehumbert, Tabor): models that, unlike classical generative linguistics, come with a relatively worked-out model of how learning is supposed to work.
- Horning (1969): unlikely classical generative grammars, statistical grammars *are* learnable without negative evidence.
- Bayesian models of learning

Bayesian models of learning: prior and likelihood

empiricist: Stolcke 1994, deMarcken 1997. Empiricist Bayesian models of grammar induction

prior: measures minimal redundancy of grammar
likelihood: how well a grammar fits input sentences

rationalist: Briscoe (1999). Parameter-setting model

prior: product of the probabilities of all the parameters
likelihood: how well a grammar fits input sentences

in between: Zuraw (2000). Loan phonology: uses Bayes Rule for listener to choose whether new word was lexical or synthesized for speaker

in between: Gildea and Jurafsky (1996), Schone and Jurafsky (2001): phonological and morphological learning as combination of (rationalist) bias and (empiricist) induction.

Objection Number 317:

OK, sure, I suppose there's a use for probabilities in language comprehension, language production, language learning, phonetics, and morphological productivity. And, OK, sure, language change and sociolinguistics have always used probabilities. And all right, I see that Pierrehumbert and Boersma are going to talk about phonology. So how about, um, pragmatics! I bet there's no use for probabilities in pragmatics!

Probability in Pragmatics

- Searle 1975 model of indirect speech act interpretation (based on Gordon and Lakoff 1971) and on (Gazdar 1981; Levinson 1983) Literal Meaning/Force Hypothesis.
- A sentence like *Can you pass the salt?* has unambiguous literal meaning of question: *Do you have the ability to pass me the salt?*
- The request speech act *Pass me the salt* is inferred by the hearer in a later step of understanding after processing the literal question.
- Many problems with Literal Force Hypothesis
- Alternative Bayesian model (Jurafsky 2003): the surface form of the sentence contains a set of probabilistic cues to the speaker's intentions. Figuring out these intentions requires inference, but not of the type that chains through literal meanings.

Objection Number 218:

You probabilistic people don’t believe in linguistic structure!

Some do, some don't!

Stored structured exemplar models: (Pierrehumbert, Bod) structure may be identical to classical models (metrical theory, syllabification), and probability is key for processing

Parameterized structure models: (Boersma) Exemplars are used to set parameters of probabilistic system (GLA); structure is identical to classical models, but acts probabilistically. Specific exemplars are not stored.

Emergent structure models (Elman, Tabor) Exemplars are used to set weights or parameters of connectionist or dynamical systems. Structure is emergent.