# Automaton Theories of Human Sentence Comprehension - Ch. 7: Information-Theoretical Complexity Metrics

2020. 03. 18

Won Ik Cho

SEOUL NATIONAL UNIVERSITY

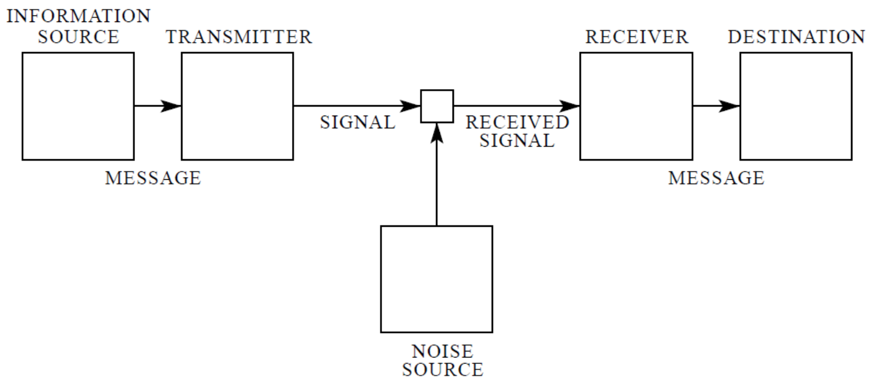
Human Interface Laboratory

# Previous approaches

- **Experience might guide a parsing mechanism**
  - Reinforcement learning / Informed search
    - Associated with an experience-based estimate of distance from completion
    - Relatively concrete
    - Could be put in correspondence with observed difficulty measures
  - But how about other direction?
    - e.g., Starting with an **abstract characterization of difficulty** itself
      and Building a **mechanical model** consistent with the characterization?

# Information theory

- **Information theory in sense of Shannon (1948)**
  - Shannon, C. E., 1948, A Mathematical Theory of Communication
    - Mathematical formulation on information, channel, transmission, receiving, noise, encoding and decoding
    - Information theoretical interpretation of 'entropy'




    - About self-information (wiki!)
      - An event with probability 100% is perfectly unsurprising and yields no information
      - The less probable an event is, the more surprising it is and the more information it yields
      - If two independent events are measured separately, the total amount of information is the sum of the self-informations of the individual events

# Probabilistic grammar

- **Information theory in sense of Shannon (1948)**
  - How we adopt this here?
    - Surprisal (Hale 2001; Levy 2008)
    - Entropy reduction (Hale 2003, 2006; Yun et al., 2015)
  - Here, historical backgrounds are managed
- **Relationship with probabilistic grammar (PG)**
  - Information theory = Logarithmic probability theory?
    - Not just a jest!
    - Important to have a sense of how probability can apply to generative grammar
  - Basic idea
    - Extend formal grammars so that the 'objects' they derive are metaphorically 'weighted'
      - Weight is typically a number
      - If these weights satisfy the axioms of probability theory?

# Probabilistic grammar

- **Relationship with probabilistic grammar (PG)**
  - Basic idea (cont'd)
    - What if the weights of derived strings add up to 1.0?
  - Suppes (1970)
    - Weighted formal languages could be viewed as **hypotheses about the distribution of utterances** in a real human language
    - … connects generative grammar to the quantitative linguistic tradition
  - How to define a probabilistic grammar?
    - Simplest one: To augment each rewriting rule with a probability

      >> ratios that have been directly off the TreeBank (PTB)
      - e.g., for NP -> DT NN,
        - » denominator: # NP appeared
        - » numerator: # NP comes with daughters as DT and NN

| | | | |
|---|---|---|---|
| 1 / 1 | S | $\rightarrow$ | NP VP |
| 160730 / 162198 | NP | $\rightarrow$ | DT NN |
| 1468 / 162198 | NP | $\rightarrow$ | NP VP |
| 1 / 1 | PP | $\rightarrow$ | IN NP |
| 3345 / 5091 | VP | $\rightarrow$ | VBD PP |
| 888 / 5091 | VP | $\rightarrow$ | VBN PP |
| 858 / 5091 | VP | $\rightarrow$ | VBD |
| 1 / 1 | DT | $\rightarrow$ | the |
| 1 / 2 | NN | $\rightarrow$ | horse |
| 1 / 2 | NN | $\rightarrow$ | barn |
| 1 / 2 | VBD | $\rightarrow$ | fell |
| 1 / 2 | VBD | $\rightarrow$ | raced |
| 1 / 1 | VBN | $\rightarrow$ | raced |
| 1 / 1 | IN | $\rightarrow$ | past |

FIGURE 34 A probabilistic grammar

# Probabilistic grammar

- **Relationship with probabilistic grammar (PG)**
  - Probability of any given derivation on a probabilistic grammar
    - … is simply the product of the probabilities of all the rules that were applied in that derivation
    - If sentence ambiguous
      >> Grammar assigns more than one derivation
      >> Total probability of the sentence is the sum of prob.s of all the derivations
  - But…
    - Preceding discussion casts PG as things that assigns prob.s to derivations
    - For a given sentence, defined:
      – Whether or not there exist any derivations for that sentence
      – If there are, what prob. goes with each of those derivations
    - How about extending this to the case of initial sentence fragments?

# Probabilistic grammar

- **Relationship with probabilistic grammar (PG)**
  - Initial sentence fragments
    - Of interest:
      - Which derivations are compatible with a sequence of words that begin
    - Not necessarily
      - What ends a well-formed sentence
    - The weights assigned by PG encode a set of grammar-based expectations about anticipated words
      - e.g., the situation by enumerating derivations that are consistent with the initial substring "the horse raced past the barn" (already been heard)
      - Longer derivations involve more rules
        - » More multipl.s by numbers <1
      - But still, overwhelming expectation at word \`barn' for the sentence to be over
      - Grammatically right < lowest!

| rank | $P(\text{derivation} \mid \text{prefix})$ | unlabelled tree |
|---|---|---|
| 1 | 0.080650290845 | [[the horse] [raced [past [the barn]]]] |
| 2 | 0.0428206028522 | [[the horse] [raced [past [the barn]]]] |
| 3 | 0.00011881492066 | [[the horse] [raced [past [[the barn] [fell [past [the horse]]]]]]] |
| 4 | 0.00011881492066 | [[the horse] [raced [past [[the barn] [fell [past [the barn]]]]]]] |
| 5 | 0.00011881492066 | [[the horse] [raced [past [[the barn] [raced [past [the barn]]]]]]] |
| 6 | 0.00011881492066 | [[the horse] [raced [past [[the barn] [raced [past [the horse]]]]]]] |
| 7 | 0.00011881492066 | [[[the horse] [raced [past [the barn]]]] [fell [past [the horse]]]] |
| 8 | 0.00011881492066 | [[[the horse] [raced [past [the barn]]]] [fell [past [the barn]]]] |
| 9 | 0.00011881492066 | [[[the horse] [raced [past [the barn]]]] [raced [past [the barn]]]] |
| 10 | 0.00011881492066 | [[[the horse] [raced [past [the barn]]]] [raced [past [the horse]]]] |
| 11 | $6.30837964401 \times 10^{-5}$ | [[[the horse] [raced [past [the barn]]]] [raced [past [the horse]]]] |
| 12 | $6.30837964401 \times 10^{-5}$ | [[[the horse] [raced [past [the barn]]]] [raced [past [the barn]]]] |
| 13 | $6.30837964401 \times 10^{-5}$ | [[[the horse] [raced [past [the barn]]]] [raced [past [the horse]]]] |
| 14 | $6.30837964401 \times 10^{-5}$ | [[[the horse] [raced [past [the barn]]]] [raced [past [the barn]]]] |
| 15 | $6.30837964401 \times 10^{-5}$ | [[[the horse] [raced [past [the barn]]]] [fell [past [the barn]]]] |
| 16 | $6.30837964401 \times 10^{-5}$ | [[[the horse] [raced [past [the barn]]]] [fell [past [the horse]]]] |
| 17 | $6.30837964401 \times 10^{-5}$ | [[the horse] [raced [past [[the barn] [raced [past [the horse]]]]]]] |
| 18 | $6.30837964401 \times 10^{-5}$ | [[the horse] [raced [past [[the barn] [raced [past [the barn]]]]]]] |
| 19 | $6.30837964401 \times 10^{-5}$ | [[the horse] [raced [past [[the barn] [fell [past [the horse]]]]]]] |
| 20 | $6.30837964401 \times 10^{-5}$ | [[the horse] [raced [past [[the barn] [fell [past [the barn]]]]]]] |
| 21 | $6.30837964401 \times 10^{-5}$ | [[the horse] [raced [past [[the barn] [raced [past [the barn]]]]]]] |
| 22 | $6.30837964401 \times 10^{-5}$ | [[the horse] [raced [past [[the barn] [raced [past [the horse]]]]]]] |
| 23 | $6.15092871436 \times 10^{-5}$ | [[the horse] [raced [past [[the barn] fell]]]] |
| 24 | $6.15092871436 \times 10^{-5}$ | [[the horse] [raced [past [[the barn] raced]]]] |
| 25 | $6.15092871436 \times 10^{-5}$ | [[[the horse] [raced [past [the barn]]]] fell] |
| 26 | $6.15092871436 \times 10^{-5}$ | [[[the horse] [raced [past [the barn]]]] raced] |
| 27 | $3.3493818379 \times 10^{-5}$ | [[the horse] [raced [past [[the barn] [raced [past [the horse]]]]]]] |
| 28 | $3.3493818379 \times 10^{-5}$ | [[the horse] [raced [past [[the barn] [raced [past [the barn]]]]]]] |
| 29 | $3.3493818379 \times 10^{-5}$ | [[[the horse] [raced [past [the barn]]]] [raced [past [the barn]]]] |
| 30 | $3.3493818379 \times 10^{-5}$ | [[[the horse] [raced [past [the barn]]]] [raced [past [the horse]]]] |
| 31 | $3.26578457301 \times 10^{-5}$ | [[the horse] [raced [past [[the barn] fell]]]] |
| 32 | $3.26578457301 \times 10^{-5}$ | [[the horse] [raced [past [[the barn] raced]]]] |
| 33 | $3.26578457301 \times 10^{-5}$ | [[[the horse] [raced [past [the barn]]]] raced] |
| 34 | $3.26578457301 \times 10^{-5}$ | [[[the horse] [raced [past [the barn]]]] fell] |

FIGURE 35 Conditional probabilities of analyses spanning the first six words

# Conditional distribution

- **Surprisal and entropy reduction**
  - Both involve summaries of conditional distributions
  - Deal with the ways that distribution changes from word to word as initial substring is lengthened
  - Intuition:
    - If this distribution changes drastically, then more information-processing work is required
      - Ought to be reflected in observable measures of sentence-processing effort
      - Surprisal / entropy : all the metric terms!
      - For $x$ an event,
        - » Surprisal: $\log_2(\frac{1}{P(x)})$
        - » Entropy: $H(x) = -\sum_i P(x_i) \log_2 P(x_i)$
          = expectation of the surprisal

# Conditional distribution

- **Surprisal**
  - Logarithm of the reciprocal of a probability
    - $\log_2(\frac{1}{P(x)})$
  - Counted in bits
  - Sometimes called as a 'self-information' of an event
    - Information value of observing 'this' outcome rather than any of the others that were possible in some predefined universe of events
    - In sentence processing, relevant event = observation of a particular successor word
  - Question of drastic vs. non-drastic change
    - Can be explained using the auxiliary concept of prefix probability
      - Prefix = initial substring
      - Nothing to do with morphology but rather comes from formal language theory (derived objects > 'words')

# Conditional distribution

- **Entropy reduction**
  - Under surprisal, processing effort is predicted to the amount of prefix probability that gets 'lost' at the transition from word to word
    - Does not matter how this is distributed
    - We only need the total amount!
  - Entropy asks whether or not the conditional distribution has gotten more or less organized since the last word
    - $H(x) = -\sum_i P(x_i) \log_2 P(x_i)$
      - Expectation of surprisal
      - $x_i$ : syntactic derivations
      - Hearer is conceptualized as trying to guess the value $X = \ x$ of r.v. representing the intended derivation of the words that are observed
      - This transit uncertainty level >> 'average' surprisal

# Surprisal and entropy reduction

- **Empirical support**
  - Hale (2001)
    - Garden path sentences and different types of relative clauses
    - Later…
      - Broad coverage of sentences
      - Eye-tracking data and neural signals (from magnetic resonance imaging)
    - Two general reasons for the productivity of the research
      - Combinability of information-theoretical complexity metrics with **essentially any model of language**
        - » Frank (2013) – RNN
        - » Park and Brew (2006) and Levy (2013) – Surprisal values from finite-state Markov models
        - » Hale (2006) and Yun et al. (2015) – Formalization of minimalism
        - » Sometimes, the data and complexity metric are not enough to decide btw alternative linguistic proposals!
      - The fact that frequency effects are among the most robust in all of psycholinguistics
        - » Surprisal and entropy - based on conditional distributions thus can **capture some syntactic effects given that the probabililistic model is grammar-based**

# Surprisal and entropy reduction

- **Difference**
  - Entropy reduction
    - Motivated by the failure of surprisal
      - In combination with context-free phrase structure grammars
    - Hale (2003):
      - Account as a more empirically-adequate replacement for surprisal
      - Builds on the different potential for recursive modification across subject- and object-extracted relative clauses
    - But whether the **full range of phenomena can be subsumed in one theory** still remains open!

Seoul National University

# EndOfPresentation

## Thank you!