Noisy-channel models in sentence comprehension

Klinton Bicknell & Roger Levy

ESSLLI 2014
Tübingen
Overview of the day

- A puzzle for surprisal theory

- Theory to handle the puzzle by revising some foundational assumptions about sentence comprehension

- Empirical study on syntactic comprehension arising from the new theory

- Using the theory to construct a rational model for eye movement control in reading and present evidence for it
Puzzle: incremental inference

- Try to understand this sentence:
  (a) *The coach smiled at the player tossed the frisbee.*

  …and contrast this with:
  (b) *The coach smiled at the player thrown the frisbee.*
  (c) *The coach smiled at the player who was thrown the frisbee.*
  (d) *The coach smiled at the player who was tossed the frisbee.*

- Readers boggle at “tossed” in (a), but not in (b-d)

---

Tabor et al. (2004, JML)
Why is *tossed/thrown* interesting?

- As with classic garden-paths, part-of-speech ambiguity leads to misinterpretation
  - *The horse* *raced* past the barn…*fell*
    - verb?
    - participle?

- But now context “should” rule out the garden path:
  - *The coach smiled at the player* *tossed*…

- A challenge for rational models: *failure to condition on relevant context*
Uncertain input in language comprehension

- State of the art models for ambiguity resolution ≈ probabilistic incremental parsing
- Simplifying assumption:
  - Input is clean and perfectly-formed
  - No uncertainty about input is admitted
- Intuitively seems patently wrong...
  - We sometimes misread things
  - We can also proofread
- Leads to two questions:
  1. What might a model of sentence comprehension under uncertain input look like?
  2. What interesting consequences might such a model have?
A first-cut answer

1. What might a model of sentence comprehension under uncertain input look like?
2. What interesting consequences might such a model have?

• **First**: a simple noisy-channel model of rational sentence comprehension under uncertain input

• **We use** probabilistic context-free grammars (PCFGs) and weighted finite-state automata (WFSA) to instantiate the model

• **Then**: we show how the introduction of input uncertainty solves the puzzle
The noisy-channel model

- Say we use a weighted generative grammar $G$ to parse a sentence $w$. We get a posterior over structures $T$:

$$P_G(T|w) = \frac{P(T, w)}{P(w)} \propto P(T, w)$$

- If we don’t observe a sentence but only a noisy input $I$:

$$P_G(T|I) \propto \sum_w P(I|T, w)P(w|T)P(T)$$

- Posterior over possible sentences:

$$P_G(w|I) \propto \sum_T P(I|T, w)P(w|T)P(T)$$

Levy (2008, EMNLP)
The noisy-channel model (II)

- This much is familiar from the parsing of speech (Hall & Johnson, 2003, 2004; Johnson & Charniak, 2004)
- Alternative scenario: we know the true sentence $w^*$ but not observed input $I$ (e.g., the study of reading)

*Expected inferences of the comprehender* marginalize over the input $I$:

$$P(w|w^*) = \int_I P_C(w|I, w^*) P_T(I|w^*) \, dI$$

$\propto Q(w, w^*)$
Representing noisy input

- How can we represent the type of noisy input generated by a word sequence?
- *Probabilistic finite-state automata* (pFSAs; Mohri, 1997) are a good model

```
vocab = a, b, c, d, e, f
```

```
Input symbol

Log-probability (surprisal)
```

“Word 1 is a or b, and I have no info about Word 2”
Probabilistic Linguistic Knowledge

- A generative probabilistic grammar determines beliefs about which strings are likely to be seen
  - Probabilistic Context-Free Grammars (PCFGs; Booth, 1969)
  - Probabilistic Minimalist Grammars (Hale, 2006)
  - Probabilistic Finite-State Grammars (Mohri, 1997; Crocker & Brants 2000)

- In position 1, \{a,b,c,d\} equally likely; but in position 2:
  - \{a,b\} are usually followed by e, occasionally by f
  - \{c,d\} are usually followed by f, occasionally by e
Combining grammar & uncertain input

• Bayes’ Rule says that the evidence and the prior should be combined (multiplied)
• For probabilistic grammars, this combination is the formal operation of weighted intersection

Grammar affects beliefs about the future
Revising beliefs about the past

- When we’re uncertain about the future, grammar + partial input can affect beliefs about what will happen.
- With uncertainty of the past, grammar + future input can affect beliefs about what has already happened.
word 1
\{b,c\} {?}

grammar

words 1 + 2
\{b,c\} \{f,e\}
The noisy-channel model (FINAL)

\[ P(w | w^*) \propto P_C(w) Q(w, w^*) \]

- For \( Q(w, w^*) \): a WFSA based on Levenshtein distance between words (\( K_{LD} \)):

Result of \( K_{LD} \) applied to \( w^* = a \ cat \ sat \)

Cost(a cat sat) = 0

Cost(sat a sat cat) = 8

\( K_{LD} \) applied to \( w^* = a \ cat \ sat \)
Near-neighbors make the “incorrect” analysis “correct”:

- The coach smiled at the player \textit{tossed} the frisbee

Hypothesis: the boggle at “tossed” involves \textit{what the comprehender wonders whether she might have seen}

Any of these changes makes \textit{tossed} a main verb!!!
The core of the intuition

- Grammar & input come together to determine two possible “paths” through the partial sentence: (line thickness ≈ probability)

  - tossed is more likely to happen along the bottom path
    - This creates a large shift in belief in the tossed condition

  - thrown is very unlikely to happen along the bottom path
    - As a result, there is no corresponding shift in belief
Ingredients for the model

\[ P(w|w^*) \propto P_C(w)Q(w, w^*) \]

- \( Q(w, w^*) \) comes from \( K_{LD} \) (with minor changes)
- \( P_C(w) \) comes from a probabilistic grammar (this time finite-state)
- We need one more ingredient:
  - a **quantified signal** of the alarm induced by word \( w_i \) about changes in beliefs about the past
Quantifying alarm about the past

- **Relative Entropy** (KL-divergence) is a natural metric of change in a probability distrib. (Levy, 2008; Itti & Baldi, 2005)
- Our distribution of interest is *probabilities over the previous words in the sentence*
- Call this distribution $P_i(w_{[0,j)})$
- The change induced by $w_i$ is the **error identification signal** $EIS_i$, defined as

\[
D \left( P_i \left( w_{[0,i)} \right) \middle\| P_{i-1} \left( w_{[0,i)} \right) \right)
\]

- new distribution
- old distribution
Error identification signal: example

- Measuring change in beliefs about the past:

  \{a,b\} \{?, \}

  \text{No change: } EIS_2 = 0 \quad \{a,b\} \{f,e\}

  \begin{align*}
  \sum_{w_1} P_2(w_1) \log \frac{P_2(w_1)}{P_1(w_1)} &= w_1=b \left(1 - \frac{1.82}{1.82}\right) + w_1=c \left(1 - \frac{0.48}{0.48}\right) \\
  &= 0.28 (1 - 1.82) + 0.72 (1 - 0.48) = 0.14
  \end{align*}

  \{b,c\} \{?, \}

  \text{Change: } EIS_2 = 0.14 \quad \{b,c\} \{f,e\}

  \begin{align*}
  \sum_{w_1} P_2(w_1) \log \frac{P_2(w_1)}{P_1(w_1)} &= w_1=b \left(1 - \frac{1.82}{1.82}\right) + w_1=c \left(1 - \frac{0.48}{0.48}\right) \\
  &= 0.28 (1 - 1.82) + 0.72 (1 - 0.48) = 0.14
  \end{align*}
Results on local-coherence sentences

- Locally coherent:  The coach smiled at the player *tossed* the frisbee
- Locally incoherent: The coach smiled at the player *thrown* the frisbee

(All sentences of Tabor et al. 2004 with lexical coverage in model)
Novel predictions of the model

• Theoretical recap:
  • Comprehension inferences involve trade-offs between *uncertain perception* and *prior grammatical expectations*
  • We saw how model may account for Tabor et al.'s (2004) local coherence sentences

• Novel prediction:
  • Uncertain-input effects should be *dependent on the perceptual neighborhood* of the sentence
Prediction: neighborhood manipulation

- Uncertain-input effects should be dependent on the perceptual neighborhood of the sentence.
- Resulting novel prediction: changing neighborhood of the context can affect EIS & thus comprehension behavior.

*The coach smiled at the player tossed the frisbee*

\(\text{that?)}\) \(\text{who?}\)

*The coach smiled toward the player tossed the frisbee*

- Substituting toward for at should reduce the EIS.

*(Levy, Bicknell, Slattery, & Rayner, 2009, PNAS)*
Model predictions

(The coach smiled at/toward the player tossed/thrown the frisbee)
There are advantages and disadvantages of both electronic and hardcopy journals. Hardcopy journals are more easily browsed, more portable and, of course people are very much used to their format. Electronic journals save on paper and their format has improved considerably over the past few years, but there are still problems over managing copyright restrictions and persuading people to use electronic instead of hardcopy journals. There is also the problem of portability. More and more journals are now being published in electronic format, although some publishers will only let you subscribe to an electronic journal provided you also subscribe to the hardcopy (more money for the same thing). Some electronic journals cost over 100% more than their equivalent hardcopy. With all these factors in mind I have been discussing individual and shared-subscriptions with the Biochemistry Department, the RSL and Blackwell’s. Whilst I feel that a move from hardcopy to electronic journals will be a very slow process in the ULP Library, electronic publishing is being carefully monitored and I would hope to introduce a few electronic texts into the Library alongside the journals which are already available for free over the Internet.
Experimental design

- In a free-reading eye-tracking study, we crossed *at/toward* with *tossed/thrown*:

  The coach smiled *at* the player *tossed* the frisbee
  The coach smiled *at* the player *thrown* the frisbee
  The coach smiled *toward* the player *tossed* the frisbee
  The coach smiled *toward* the player *thrown* the frisbee

- Prediction: interaction between preposition & ambiguity in some subset of:
  - Early-measure RTs at critical region *tossed/thrown*
  - First-pass regressions out of critical region
  - Go-past time for critical region
  - Regressions into *at/toward*
Experimental results

The coach smiled at the player tossed...

---

First-pass RT  
Regressions out  
Go-past RT  
Go-past regressions  
Comprehension accuracy

---

Noise level (low=noisy)
great evidence for noisy channel comprehension

• readers maintain uncertainty about the past
• readers update that uncertainty on the basis of new input
• something like EIS seems to explain (at least part of) Tabor et al.'s (2004) local coherences effects
Surprisal and EIS

processing difficulty and eye movements

- two types of processing difficulty: surprisal and EIS
- unclear *how* these types of difficulty affect eye movements?
  - more fixations? longer fixations? regressions?
- unclear *why* these types of difficulty affect eye movements
  - why keep looking at a word in proportion to surprisal?
  - why make regressions when EIS is high?

next up: a principled view of reading that answers these questions: reading as efficient visual information gathering
A new view of reading

- rich linguistic knowledge and perceptual input combine to identify text

- trade-off between prior probability and visual input: when linguistic context provides more information, need less perceptual input to be confident
A new view of reading

• eye movement decisions (how long to fixate, where to move next) made to get visual input (cf. Legge et al., 1997)

• get more perceptual information about a word by fixating it longer/more

• language system determines the most useful visual input to get
  -> when/where to move the eyes

  • reading as active

  • principled solution: optimal control, decision theory
How would this link yield interesting linguistic effects?

trade-off between prior probability and visual input

efficient language users will make fewer and shorter fixations on words with high probability given context

efficient language users will be more likely to skip over words with high probability given context

look at quantitative predictions of this account for shape of linguistic effects via a computational model
A computational model

- probabilistic inference
- efficient eye movement control
- realistic environment

Bicknell & Levy (2010, 2012)
A computational model

probabilistic inference

• combines visual information with probabilistic language knowledge using methods from computational linguistics

visual information as weighted finite-state automata (wFSAs)

efficient composition

probabilistic language knowledge

PCFG:
S -> NP VP / 0.9
S -> S Conj S / 0.05

n-grams:
p(York|New) = .4
p(Brunswick|New) = .01

efficient inference: closed under intersection with wFSAs

Bicknell & Levy (2010, 2012)
A computational model

**efficient eye movement control**

- determines efficient eye movement behavior in response to incremental comprehension using machine learning methods

- to maximize reward $R$, a linear function of speed and accuracy
  - first-cut accuracy: log probability of sentence string under model beliefs after reading

- *parameterized* behavior policies control the eyes

- determine parameters that maximize $R$ using reinforcement learning (PEGASUS algorithm, Ng & Jordan, 2000)

  fit model behavior to maximize reward, not fit

  Bicknell & Levy (2010, 2012)
A computational model

**realistic environment**

- embed model in realistic environment using findings from psychophysics and oculomotor control
- exponentially decreasing visual acuity
- saccade initiation delays
- normally-distributed motor error

Bicknell & Levy (2010, 2012)
A computational model

probabilistic inference

efficient eye movement control

realistic environment

Bicknell & Levy (2010, 2012)
Model simulation results

**methods**

• evaluate model on human effects of word frequency and predictability

• **frequency**: word's overall rate of occurrence in language
  
  • more frequent: dog
  
  • less frequent: parsnip

• **predictability**: probability of word in context
  
  • more predictable: The children went outside to *play* …
  
  • less predictable: My friend really likes to *play* …

Bicknell & Levy (2012)
Model simulation results

methods

• simulate model reading typical psycholinguistic sentences

• analyze three word-based eye movement measures

• only fit one free parameter of model (visual input rate) to match average human word reading time. (others fit to optimize reading efficiency)

  • any match to effect shape falls out of principles of efficient identification and linguistic structure

Bicknell & Levy (2012)
Model simulation results

**frequency**

![Log frequency vs Gaze duration (ms)](image)

**predictability**

![Log predictability vs Gaze duration (ms)](image)

gaze duration: total duration of all fixations on word

Bicknell & Levy (2012)
Model simulation results

**frequency**

<table>
<thead>
<tr>
<th>Refixation probability</th>
<th>Log frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>model</td>
</tr>
</tbody>
</table>

**predictability**

<table>
<thead>
<tr>
<th>Refixation probability</th>
<th>Log predictability</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>model</td>
</tr>
</tbody>
</table>

refixation probability: probability of making more than one fixation on the word

Bicknell & Levy (2012)
Model simulation results

**frequency**

<table>
<thead>
<tr>
<th>human</th>
<th>model</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td></td>
</tr>
</tbody>
</table>

Log frequency

Skipping probability

**predictability**

<table>
<thead>
<tr>
<th>human</th>
<th>model</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td></td>
</tr>
</tbody>
</table>

Log predictability

Skipping probability

*skipping probability: probability of not directly fixating word*

Bicknell & Levy (2012)
Model simulation results

**frequency**

predicts shape of all effects

predictability

Bicknell & Levy (2012)
Part One: Summary

- implemented model efficiently controlling eyes to gather perceptual info for text identification

- model well predicts shape of linguistic effects on eye movements, fitting only one parameter to average reading time

- provides motivated reason why linguistic effects appear on eye movements in reading

  • reason is not because language processing operations take substantial time (in fact, they are instant in the model)
probabilistic inference

[perceptual input]

rich probabilistic language knowledge

likelihood

prior

beliefs about identity of the text

rational action

[comprehension behavior]

principle:
Get more input about words with low prior probability given context [cf. surprisal]
Part Two

Viewing reading as visual information gathering solves the puzzle of regressions.
The puzzle

~10% of saccades move the eyes back (‘regress’) to previous words

Why would it be useful to regress to previous words so often?

movie by Piers Cornelissen
A solution

confidence falling

‘From the closet, she pulled out a #acket for’ the upcoming match’

\[ p(\text{jacket}) = .9 \quad \rightarrow \quad p(\text{jacket}) = .4 \]
\[ p(\text{racket}) = .1 \quad \rightarrow \quad p(\text{racket}) = .6 \]
\[ p(\text{packet}) = .0 \quad \rightarrow \quad p(\text{packet}) = .0 \]

confidence high \quad \rightarrow \quad confidence low

Maybe regressions an effective way to deal with this?

Bicknell & Levy (2010)
A solution

simulations

• compare behavior policies that make regressions to those that don't

• regressive policies are strictly more efficient (both faster and more accurate)

but do humans make regressions when confidence falls?

Bicknell & Levy (2010)
Human regressions

methods

• Dundee corpus: large corpus of eye movements reading newspaper editorials [Kennedy & Pynte, 2005]

• derived broad coverage measure of confidence falling: $\Delta_c$
  
  • $\Delta_c = \log\text{ frequency} - \log\text{ predictability}$
    (ask me about the derivation!)

• logistic mixed-effects regression model predicting whether or not the eyes will make a regression to a previous word
  
  [statistics]
  
  [eyes] Bicknell & Levy (2011)
Human regressions

methods

• predictors

  • $\Delta_c$ for current word (word$_n$) & previous word (word$_{n-1}$) for late-triggered regressions

  • variables suggested by other accounts of regressions (frequency, length, landing position, saccade length, fixation duration)

• predict more regressions for high $\Delta_c$ (on either word)

Bicknell & Levy (2011)
Human regressions

Relevant for today

- more regressions when current word or previous word makes confidence fall
- confidence falling one of most reliable predictors

Bicknell & Levy (2011)
Part Two: Summary

explaining regressions

- efficient reader will update uncertain beliefs about prior material on the basis of new input: confidence falling

- confidence falling & regressions
  - regressions help make reading more efficient
  - confidence falling explains short-range regressions in naturalistic text

- can differentiate types of ‘processing difficulty’
  - words of low prior probability -> more/longer fixations
  - losing confidence in prior material -> regressions
1. Get more input about words with low prior probability given context [cf. surprisal]
2. Regress to previous words if confidence about them falls [cf. EIS]
Today's lecture argued that comprehension is well understood as probabilistic inference on noisy perceptual input.
Conclusion

what this buys us

• part 1: a solution to the problem of local coherences
  • they change beliefs about what readers thought they saw (EIS)

• part 2: a principled reason why reading times on a word should be well described by surprisal
  • need less visual information to be confident in identity of words that have high contextual probability

• part 3: a principled reason why readers should make regressions
  • language often makes readers doubt what they thought they saw
Conclusion

lots more recent work in this area (papers on class website)

- Gibson et al. (2013): evidence for noisy-channel inferences about what sentences mean
  - "The ball kicked the girl" can mean "The girl kicked the ball"
- Lewis et al. (2013): evidence that readers rationally change their eye movement control given their goal functions (as predicted by a computational model)