Commentary at Workshop on Exemplar-based Models in Linguistics

Christopher Manning
Stanford University
http://nlp.stanford.edu/~manning/

Some quotes

- “When I say, for instance, ‘I had a good breakfast this morning,’ it is clear that I am not in the throes of laborious thought, that what I have to transmit is hardly more than a pleasurable memory symbolically rendered in the grooves of habitual expression... It is somewhat as though a dynamo capable of generating enough power to run an elevator were operated almost exclusively to feed an electric doorbell.”
  - E. Sapir, Language, 1921

- “We should never underestimate the power of simply remembering some or all of the examples and comparing test examples with our memory.”
  - B.D. Ripley, Pattern Recognition and Neural Networks, 1996

- “The cost for computation is decreasing roughly by a factor of two every year, whereas the price paid for incorrect assumptions is remaining the same.”

Some quotes

- People are very sensitive to frequency when processing and producing language
  - E.g., Subcategorization probabilities of verbs play role in reading time (Clifton, Jr. et al. 1994; Trueswell et al. 1991; Jennings et al. 1997; MacDonald 1994). Basically anything that can be measured in language shows frequency sensitivity.
  - It was a mistake in early linguistics to think that limitations on memory are a major concern in models of human cognition.
  - The goal of avoiding “redundancy” in the grammar was misguided

Rens Bod: Exemplar-Based Syntax

- Bod presented data-oriented linguistics
  - A form of instance-based learning/processing

- But is it actually “exemplar-based?”
  - If “exemplar-based” means simply that you somehow make use of all the examples that you’ve ever heard, then it is.
  - If “exemplar-based” means that you have extracted “exemplars” as a layer of abstraction above examples [as Joan Bybee seemed to assume last night] then it isn’t.
  - If “exemplar-based” means that you are finding closest examples at runtime, then it isn’t: the model can be viewed as using sufficient statistics drawn from all examples (which don’t need to be stored)

Rens Bod: Exemplar-Based Syntax

- The emphasis is on finding universal representations rather than universal grammar.
  - But is the different between rules and representation exaggerated?

  - Rules
    - S → NP VP
    - S → NP[fem.sg] VP[fem.sg]
  - Representations
    - S
    - NP VP
    - NP[ fem.sg] VP[fem.sg]
Rules vs. representations

- The distinction between using a grammar and a corpus of stored linguistic representations ends up somewhat unclear.
  - The use of large fragments reminds one of example-based (KNN) approaches. However, these compute no aggregate statistics, but just reason from most similar stored examples.
  - Rather than using individual examples, TD aggregates statistics over all fragment analyses. The sufficient statistics of the corpus can be thought of as a probabilistic grammar inherent in the corpus. You don’t need to keep the examples....
  - Cf. simulation methods in statistics.

Rules vs. representations

- The interesting difference is the degree of abstraction/decomposition (the size/detail of chunks).
  - Having large tree/representation pieces can model idioms/prefabs/collocations.
  - It can model non-head dependencies, such as between a superlative and a following PP: the fastest woman in the world.
  - It can also just model any non-independence effect that occurs at a larger scale, and these are common.
    - Ex. NPs differ in distribution as subject vs. object.

Modeling reaction time via frequency

Parallel, expectation-based syntactic processing

- Hale 2001, Levy 2005 LSA model:
  - The more a word is expected, the easier it is to process; difficulty = surprisal(w).
  - SURPRISAL(w) = log Prob(w | CONTEXT).
  - Parallel parsing with a probabilistic context-free grammar (PCFG) determines the expectation of a word.
  - Very similar to TD, but use of surprisal better founded than the measure he uses?

Locality-based models for final verbs

- In German, Gibson-style locality-based models (Gibson 1998) predict difficulty for longer clauses.
  - Er hat die Gruppe geführt
  - Prediction: easy
  - Er hat die Gruppe auf den Berg geführt
  - Prediction: hard

Empirical results

- But Konieczny (2000) found that final verbs were read faster in longer clauses.
  - Er hat die Gruppe geführt
  - Prediction Result: easy slow
  - Er hat die Gruppe auf den Berg geführt
  - Prediction Result: hard fast

Predictions from Hale 2001/Levy 2004 model

- Er hat die Gruppe (auf den (sehr schönen) Berg) geführt
  - Prediction: hard fast
  - Er hat die Gruppe geführt
  - Prediction: easy slow

Memory-based models (e.g., Gibson 1998) would violate monotonicity.
Christopher Johnson

- Historical reinterpretation recapitulated in language acquisition
  - Systematic ambiguities in meaning can actually drive learning
  - Extension from source to target sign

- Johnson provides evidence that this sometimes happens
  - Here’s an example where it might not be true...

Example: kind/sort of
(Tabor 1994)

- Kind/sort can head an NP or be an adverbial modifier:
  - [That kind of knife] isn’t used much.
  - We are [kind of] hungry.

- There is a path of reanalysis through ambiguous forms a la Johnson:
  - [a kind of dense rock]
  - [a [kind of] dense rock]

Example: kind/sort of

- Jesus Christ his owne son through kind
  - OE: has kind little/no kind of:
  - A nette sent in to the see, and of alle kind of fishe gesdyng (1382 Wyeliff)
  - Their finest and best, is a kind of course red cloth (1570 True Report)
  - I was kind of provoked at the way you came up (1830 Mars. Spy)

- But, surely “kids today” learn the softener use of kind of...!
  - What is the scope of the claim?

Example: near

(Cf. Maling 1983)

- In Middle English, an adjective
  - Today is it an adjective or a preposition?
    - The near side of the moon
    - We were near the station
  - Not just a word with multiple parts of speech!
     Evidence of blending:
    - I was nearer the bus stop than the train

Gerund V-ing forms

- Classically described as an alternation between nominal and verbal gerund
- Early example of a “squish” (Ross 1972)
- Really, V-ing forms can appear to be of any of the core categories of Chomsky (1970):

<table>
<thead>
<tr>
<th>V</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>an unassuming man</td>
</tr>
<tr>
<td>-</td>
<td>he is eating dinner</td>
</tr>
</tbody>
</table>
Mixed noun/verb status

• Distinguishing N/V uses of V-ing is a standard syntax problem, but, historically, the V-ing form shows a history of increasing verbalization
• In many periods it shows a notably non-discrete status.
  - Noun deenon: appearing with a determiner
  - Verb criterion: taking a direct object
• The non observing role is the one which the world has blamed in our societies (Brodie, Essay Gram. Rev. 310, 1848)
• The only mental provision she was making for the evening of life, was the collecting and transcribing all the riddles of every sin that she could meet with, (Jane Austen, Emma, 1815)
• The difficulty is in the gaining the gold into Trewbor (Sam Butler, Trewbor Revisited, 1902)

Continued noncategoricalness

• Graded judgements continue in this domain (Horn 1975):
  - Tom's winning the election was a big upset
  - His teasing John all the time has got to stop
  - There is no marking exams on Fridays
  - The cessation hostilities was unexpected
• Houston (1985: 320) shows that an assignment of (ING) forms to a discrete POS classification is less successful [in a predictive sense] than a continuum a la Ross in explaining -ING vs. -IN
  - "grammatical categories exist along a continuum which does not exhibit sharp boundaries between the categories."

Exemplar–based models

• A better approach to these "syntactic nuts" would be an important conceptual advance
• Models that place individual words in a syntactico-semantic space based on proximity - that is, the kind of exemplar–based model examined at this workshop - seem the best hope of dealing with cases of mixed and gradient categorization

- The End -