Grammar Engineering

May 8, 2006

Precision grammars and corpora

Overview

- Theoretical motivation
- Methodology
- Results
- Your grammars
- Precision grammars and NLP

Theoretical motivation (1/2)

- Corpora as a sole source of data are inadequate because:

  They are limited in size and may not reflect the full range of grammatical constructions.

  They contain errors due to processing and reflect other extragrammatical factors.

  They can only provide positive (attested) examples, and not contrasting negative ones.

Theoretical motivation (2/2)

- Intuitions as data are inadequate because:

  Grammaticality is neither homogeneous nor categorical.

  Grammaticality judgments are frequently formed in unnatural context vacuums.

  Social/cultural biases color judgments.

  Relying solely on intuitions limits linguists to only the data they have the imagination to think up.

Combine the two types of data for better results!

- Grammar engineering provides a sophisticated way of doing so.
- Precision grammars encode a sharp notion of grammaticality.
- Use grammar as a representation of intuitions.
- Use the corpus as a source of further data to explore.
- Process the corpus with the grammar...

Methodology

- Randomly select 20,000 strings (‘sentence tokens’) from the BNC written component.
- Strip punctuation, tag for part-of-speech, tokenize proper names and number expressions, normalize to American spelling.
- Select those strings with full lexical span (32%).
- Process these strings with the ERG to isolate those that can’t presently be parsed.
- Use treebanking technology/methodology to validate parses.
- Propose paraphrases of the unparseable strings until the ERG is able to parse one.

Results: Grammar coverage

- 57% of strings parsed.
- 83% of parsed strings assigned a correct (preferred) parse, perhaps among others.
- Average ambiguity for 10-20 word strings: 64 parses.

Results: Causes of parse failure

Cause of parse failure	Frequency	Category
Missing lexical entry	41%	grammar
Missing construction	39%	grammar
Fragment	4%	grammar
Preprocessor error	4%	neither
Parser resource limitations	4%	neither
Ungrammatical string	6%	corpus
Extragrammatical string	2%	corpus

Missing lexical entries (1/2)

- Incomplete categorization of existing lexical items

  table as a verb

  'universal grinder'

- Syntactically-marked MWEs

  take off, verb + *up*

  off screen, at arm's length

  High frequency: verb-particles constitute 1.6% of BNC word tokens

Missing lexical entries (2/2)

- Drawbacks to introspection alone: subtle gaps like transitive *suffer*
- Drawbacks to corpus data alone: *tell* in the ‘discover’ sense:

  @Not sure how you can tell.

  Can/could you tell?

  Are you able to tell?

  *They might/ought to tell.

  How might you tell?

  *How ought they to tell?

Missing constructions (1/4)

- [@] *However pissed off* we might get from time to time...
- ERG specifically disallowed this.
- → Corpus data as a check on introspection.
- Further corpus investigations surprised ys.

Missing constructions (2/4)

- [@] He's a good player and a *hell of a* nice guy, too.
- Baldwin et al present this as a semantic puzzle:

  Apparent syntactic attachment to NP/N′ because of definiteness restrictions

  Semantic attachment to adjective (intensifier)
- Still complex, but less mysterious, in a world where definiteness is encoded as a feature of indices.

Missing constructions (3/4)

- [@]The price of train tickets can vary from *the reasonable* to *the ridiculous*.
- Exocentric NPs not limited to classes of people.
- What adjectives can appear here, and with what kinds of referents?

Missing constructions (4/4)

- @This sort of response was also noted in the sample task for *criterion 2*.
- ‘Title’ (common noun) + series element
- Frequent in corpora (like dates, number names, quotatives)
- Not usually remarked on in syntactic theory

Extragrammatical strings

- Prime example: Structural markup:

  @There are five of these general arrest conditions: (a) the name of…

- Preprocessing requires interface to grammar:

  @(I) The Mrs Simpson could never be Queen.

  @(I) rarely took notes during the thousands of informal conversational interviews.

Summary

- Methodology goes beyond merely using the corpus for inspiration.

  encoding intuitions in the grammar

  use the grammar to process the corpus, twice: filter out ‘easy’ cases, investigate where in a string the problems are

- Provides detailed feedback to grammar developers

- Turns up previously unnoted constructions, which might be too low frequency to be found otherwise

How about your grammars?

- Role of corpora so far?
- How to get from current state to something that could turn up unexpected constructions?

Precision grammars in NLP

- Baldwin et al: Notion of grammaticality cuts down on spurious ambiguity and crucial in avoiding ill-formed output in generation
- Elsewhere: Value of elaborated semantic representations
- Cost: Could grammar development ever become cheaper than treebank development?

Overview

- Theoretical motivation
- Methodology
- Results
- Your grammars
- Precision grammars and NLP