Integrating WordNet with NL-Soar

Deryle Lonsdale

Department of Linguistics
Brigham Young University
Provo, UT 84602
(801) 378-4067
lonz@byu.edu

NL-Soar review

- natural-language modeling application of Soar
- various language tasks: comprehension, generation, mapping, discourse
- various task integrations: NTD-Soar, TacAir-Soar, AML-Soar, SI-Soar, ESL-Soar

- buffered, decay-susceptible input/output
- u(tterance)-model: X-bar syntax piecewise via u-cstr op's
- s(ituation)-model: LCS semantics piecewise via s-cstr op's
- lexical access: word syn/sem features via access op's

- comprehension: sentence $\mapsto$ u-model $\mapsto$ s-model
- generation: s-model $\mapsto$ u-model $\mapsto$ sentence

- operators learnable, interleaved, shared
- English, French, Polish implemented to various degrees

WordNet overview

- developed by Princeton cogsci researchers
- large-scale lexical database for English
  - comprehensive, systematic lexical inventory
  - lexical word senses
  - lexical subcategorization
  - lexical hierarchy
  - lexical morphology interface
- public domain, free
- widely used in NLP and by psycholinguists
- more information: `www.cogsci.princeton.edu/~wn`

WordNet example

```
The noun "dog" has 6 senses in WordNet.

1. dog, domestic dog, Canis familiaris
2. frump, dog -- (a dull unattractive unpleasant girl or woman)
3. dog -- (informal term for a man: "you lucky dog")
4. cad, bounder, blackguard, dog, hound, heel
5. pawl, detent, click, dog -- (a hinged device...of a ratchet)
6. andiron, firedog, dog, dogiron -- (metal supports for logs in a fireplace)


The verb "dog" has 1 sense in WordNet.

1. chase, chase after, trail, tail, tag, dog, go after, track


==================================

WordNet 1.6 results for "Hypernyms (this is a kind of...)" search of noun
"dog"

6 senses of dog

Sense 1

dog, domestic dog, Canis familiaris
 => canine, canid
   => carnivore
      => placental, placental mammal, eutherian, eutherian mammal
          => mammal
              => vertebrate, craniate
                  => chordate
                      => animal, animate being, beast, brute, creature, fauna
                          => life form, organism, being, living thing
                              => entity, something
```

Why WordNet in NL-Soar?

- NL-Soar lexicon shortcomings
  - coverage: very few words (a few hundred)
  - arbitrary word sense distinctions
  - syn/sem feature inconsistencies
  - inflected forms required
  - no hierarchical relations
- widely used in NLP tasks, corpora
  - inferencing
  - summarization
  - parsing
  - word-sense disambiguation
  - psycholinguistic modeling
- extension to other languages

The approach

- separate image from canonical NL-Soar
- lexicon productions replaced with generic access functionality
- supplements (so far, doesn't replace) old information with new lexical data
- folded into lexical access operator
- API, but desired features not all supported
- TCL code ($\approx$ 2k lines) to traverse WordNet, extract needed information
- few iterations of code optimization

Providing morphology

- WordNet interface called Morphy
- first integration of morphology into NL-Soar
- returns base form(s) for inflected words
- problem: doesn't return features
- TCL code to induce features by reverse-engineering results returned from Morphy

- before:
  - coverage was a few hundred words
  - no on-the-fly morphology; all hard-coded in lexicon
- now:
  - over 94,000 words
  - full morphological reduction, all possibilities pursued

Providing syntax

- subcategorization frames: what categories of words/phrases are allowed together

```
vframe    subcat
------    ------
 2 0      We ask.
 8 NP     We ask a question.
 9 NP     We ask the man.
11 NP     The books ask a question.
14 NP NP  We ask him a question.
16 NP PP  We ask a favor from him.
20 NP PP  We ask him for a favor.
22 PP     We ask for a favor.
24 NP TO  We ask him to call.
26 THAT   We ask that he calls.
28 TO     We ask to call.
29 WH-INF We ask whether to call.
```

- dynamically calculates possible frames, matches with current syntactic environment
- much more versatility for argument dropping, adding, etc.

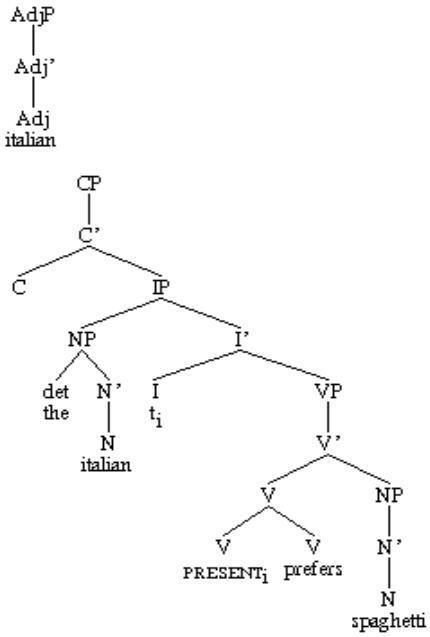
Providing semantics

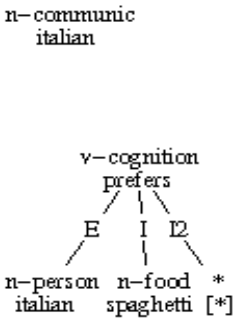
- primitive senses: based on high-level WordNet sense categories (e.g. n_animal, v_body)
- more explicit representation of semantic ambiguity (e.g. dog is an n_animal and an $n\_artifact$)
- little or no preposition, adjective, adverb semantics

- before: handful of semantic primitives (e.g. thing, path, action) only vaguely separable
- now: pre-defined inventory of senses

Sample structures

The Italian prefers spaghetti.





Issues

- much more frequent ambiguity (morphological, syntactic, and semantic)
- category limit of two is too small for syntax, semantics (but three seems just right) (e.g. "He flies planes.")
- more backtracking to undo wrong hypotheses, so much more versatile snip operator required
- performance: too slow, so load in meta-index to WordNet when initializing NL-Soar
- memory usage: often runs out in long or massively ambiguous sentences

Current status

- exercising the integration
- syntax: match previous non-WordNet baseline
- semantics: see Rytting (next talk)
- learning: chunked up as part of lexical access operator

Future work

- leverage the hypernyms
- upgrade baseline NL-Soar to Soar8, newer TCL
- package baseline NL-Soar
- package, release N(W)L-Soar
- learning from sense-tagged corpora
- EuroWordNet for other languages
- re-implement TCL interface in C (?)

Conclusions

- coals
  - leveraging off-the-shelf resource
  - level of analysis (dealing with fine granularity)
  - TCL is fairly slow
  - big leap from little information to copious amounts
- nuggets:
  - coverage (scaling up in a big way)
  - principled approach to senses, disambiguation
  - foundation for semantics-based processing
  - compatibility with other NLP resources
  - groundwork for modeling experimental results