Analogical Modeling of Language in Soar

Deryle Lonsdale

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

(NL-)Soar and language modeling

- syntax
- semantics
- discourse

BUT:

- no phonology, morphology, or related areas
- minimal lexical acquisition, selection

Analogical Modeling of Language

- active research paradigm in language modeling
- data-driven exemplar-based approach
- no explicit encoding of rules
- non-rule-based, non-connectionist architecture
- outcome prediction based on contextual parameters
- statistical method derives analogical set
- more versatile than nearest-neighbor approaches
- robustness: novel data, noisy data

AML research (language-specific)

- Finnish past tense (Skousen 1989)
- German plurals (Wulf 1996)
- Dutch stress (Daelemans et al. 1994)
- Arabic lexical selection (Parkinson)
- Japanese loanword formation (Blaylock)
- Spanish gender, stress (Eddington 1998)
- Turkish morphophonemics (Rytting)
- English past tense (Chandler 1997)
- English negative prefixes (Baltes et al.)
- Chinese classifiers (Bourgerie)

AML research (linguistics-related)

- statistical language modeling (Skousen 1998)
- psycholinguistics (Derwing & Skousen 1989)
- comparison to connectionist, dual-route models (Chandler, Eddington)
- machine translation (Jones 1996)
- NLP applications (Lonsdale 1999)
- international conference at BYU in March 2000
- web site: http://humanities.byu.edu/aml/homepage.html

Why AML in (NL-)Soar?

- complementarity with existing functionality
  - low-level linguistic functions for Soar
  - high-level functions for AML
- instance-based learning for language in Soar
- testbed for modeling (future) AML psycholinguistic results
- integration with cognitive processes, other tasks
- performance issues

Running the AML system

- (encoded) set of data items
- set of possible outcomes
- system dynamically processes dependencies, relationships
- probable outcomes based on observed data and derived generalizations
- produces statistical results (w/rt outcomes)
- shows contribution of data instances (analogical set)

Data instance encoding

- feature vector representing salient properties of data instances
- vector length is constant across data set
- nondeterministic mappings possible
- Finnish verb sample data:

```
A HEVI0=OTTa HEITTa
A HIVI0=OHTa HIIHTa
A HO0=0=OHTA HOHTA
A HOVI0=0=TA HOITA
C HUVOSL0=TA HUOLTA
C HUVU0=0=TA HUUTA
A II0=0=0=ME IME
A II0=0=OSKE ISKE
A II0=0=OTKE ITKE
A II0=0=0=Ta ITa
```

Soar implementation

- read data items (one operator)
- read test item (one operator)
- set up requisite data structures (two operators)
- calculate lattice of contextual dependencies
  (n features: $n + 1$ operators)
- compute frequencies (one operator)

Sample trace

```
amlsoar> r

     0: ==>S: S1
     1:    O: (amldata: )
     2:    O: (amltest: )
     3:    O: (getvec: )
     4:    O: (getlevels: )
     5:    O: (levelop: 0)
     6:    O: (levelop: 1)
     7:    O: (levelop: 2)
     8:    O: (levelop: 3)
     9:    O: (levelop: 4)
    10:    O: (levelop: 5)
    11:    O: (levelop: 6)
    12:    O: (levelop: 7)
    13:    O: (levelop: 8)
    14:    O: (levelop: 9)
    15:    O: (levelop: 10)
    16:    O: (reportfreqs: )
Outcome:A Freq: 22
Outcome:B Freq: 5
  Goal g-aml succeeded.
g-aml achieved
System halted.

amlsoar>
```

Current status

- implemented: all core functionality
- 285 productions
- Finnish verb coverage

Future work

- improve interface (file i/o)
- implement memory functionality
- more scoring functions
- optimization
- experiment with (sub-)goal structure
- explore learning
- integrate with NL-Soar