Combining Lexical Resources in a Robust Broad-Coverage Semantic Parser

John Dowding
Mathew Purver

7th Annual Semantics Fest
March 10, 2006

STANFORD
UNIVERSITY

THE UNIVERSITY OF CALIFORNIA
LET THERE BE LIGHT
SANTA CRUZ

Open-Domain Interpretation

- Extract propositional content from meetings
    - Used to help detect decisions and action items
    - Part of DARPA’s CALO program,
        - emphasizing “learning in the wild”
- Open-Domain
    - Meeting topics are not specified in advance
- Analyzing speech recognition output
    - Word Error Rates near 30%
    - Word Confusion Networks encoding large numbers of speech recognition hypotheses
        - Avg. 1.9x10^{34} paths (mean)

Approach

- *Given the prevalence of ill-formed data, allowing for the full complexity of English syntax seems likely to introduce more errors than it fixes.*
- Emphasize extracting predicate-argument structure
- Extract major phrase types (S, VP, NP, PP)
  - Rely heavily on lexicon
  - Less emphasis on grammar
- Build lexicon from publicly available resources
  - COMLEX, VerbNet, WordNet, NomLex
  - Combine semantic information across resources
- Avoid hand-modifying the lexicon

Lexical Resources

- COMLEX provides detailed syntactic features
  - 23,195 nouns (mass/count and temporality)
  - 5,665 verbs (subcategorization)
  - 4,200 adjectives (gradeability and subcategorization)
  - 3,120 adverbs (syntactic distribution)
  - Provides morphological variants for irregular forms
- VerbNet provides semantic information for 5,000 verbs
  - Verb class
  - Thematic Roles
  - Syntax-Semantics Mapping
  - Selectional Restrictions
    - Expressed as concepts from the EuroWordNet upper ontology

Lexical Resources (continued)

- WordNet
  - We take another 15,500 nouns from WordNet
  - Semantic class information for all nouns
  - Semantic classes hand-aligned to the EuroWordNet upper ontology
- NOMLEX (and NOMLEXPLUS)
  - Syntactic information for event nominalizations
  - Mapping into corresponding verb syntactic positions
  - Aligned with VerbNet to provide selection on noun arguments.
- Common proper names from US Census data

Pruning low-frequency POS

- COMLEX contains many entries for low-frequency part-of-speech assignments for high-frequency words.
    - Examples like are, down, low, okay
- These caused trouble for the parser
- Used hand-tagged data (Switchboard, ATIS, WSJ) to identify low-frequency POS assignements
    - Pruned POS when a word had a dominant POS (>98%)
- Eliminated POS assignments for ~900 words.

Minimal Recursion Semantics (MRS)

- Based on Copestake, Flickenger, Sag (1999)
- Flat semantic representation that underspecifies scope
- Identifies entities and events
- Represents elementary predications
- Easy to extract features for machine learning
  - Additional ML approaches to detecting action items

MRS Example

B:declarative(C)

D:quant(exists;[det],F;[get-13.5.1],H,I)

J:event(F;[get-13.5.1])

J:'Buy_v'(F;[get-13.5.1])

K:agent(F;[get-13.5.1],L;[organization])

K:theme(F;[get-13.5.1],N;[phys_obj])

V:quant(a;[indef],N;[phys_obj],W,X)

Y:entity(N;[phys_obj])

Y:new_adj(N;[phys_obj])

Y:computer_n(N;[phys_obj])

Z:quant(the;[def],L;[organization],A1,B1)

C:entity(L;[organization])

C1:department_n(L;[organization])

The department bought a new computer

NOMLEX Example “talk”

```
(NOM :ORTH "talk"
   :VERB "talk"
   :NOM-TYPE ((VERB-NOM))
   :VERB-SUBJ ((N-N-MOD)
         (DET-POSS)
         (PP :PVAL ("by")))
   :SUBJ-ATTRIBUTE ((NHUMAN))
   :VERB-SUBC ((NOM-INTRANS :SUBJECT ((N-N-MOD)
                     (DET-POSS)
                     (PP :PVAL ("by")))
                :REQUIRED ((SUBJECT)))
         (NOM-PP-PP :SUBJECT ((N-N-MOD)
                    (DET-POSS)
                    (PP :PVAL ("by")))
               :PVAL ("about" "of" "on")
               :PVAL2 ("to" "with"))
         (NOM-PP :SUBJECT ((N-N-MOD)
                  (DET-POSS)
                  (PP :PVAL ("by")))
               :PVAL ("about" "on" "of" "to" "with"))
```

VerbNet Thematic Roles “talk”

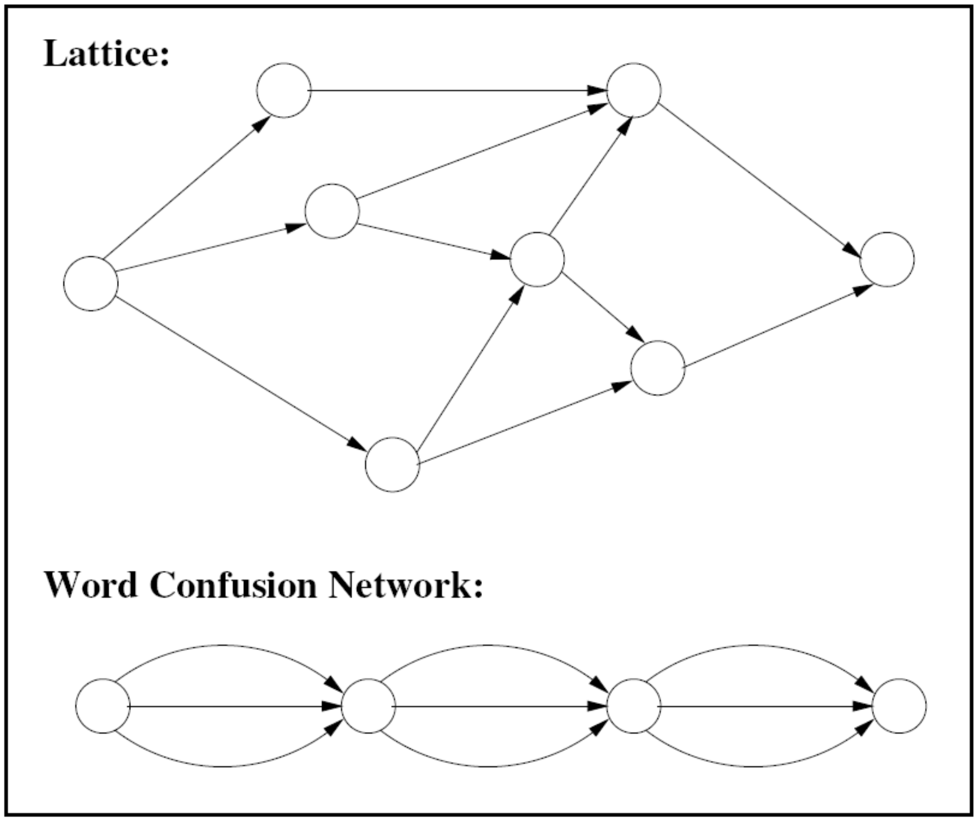
```
<THEMROLES>
    <THEMROLE type="Actor">
      <SELRESTRS logic="or">
        <SELRESTR Value="+" type="animate"/>
        <SELRESTR Value="+" type="organization"/>
      </SELRESTRS>
    </THEMROLE>
    <THEMROLE type="Actor1">
      <SELRESTRS logic="or">
        <SELRESTR Value="+" type="animate"/>
        <SELRESTR Value="+" type="organization"/>
      </SELRESTRS>
    </THEMROLE>
    <THEMROLE type="Actor2">
      <SELRESTRS logic="or">
        <SELRESTR Value="+" type="animate"/>
        <SELRESTR Value="+" type="organization"/>
      </SELRESTRS>
    </THEMROLE>
    <THEMROLE type="Topic">
      <SELRESTRS>
        <SELRESTR Value="+" type="communication"/>
      </SELRESTRS>
    </THEMROLE>
  </THEMROLES>
```

VerbNet Frame “talk”

```
<FRAME>
     <DESCRIPTION descriptionNumber="0.1" primary="PP-PP" secondary="to-PP Topic-PP" xtag=""/>
     <EXAMPLES>
       <EXAMPLE>&quot;Susan talked to Rachel about the problem&quot;</EXAMPLE>
     </EXAMPLES>
     <SYNTAX>
       <NP value="Actor1">
         <SYNRESTRS/>
       </NP>
       <VERB/>
       <PREP value="to">
         <SELRESTRS/>
       </PREP>
       <NP value="Actor2">
         <SYNRESTRS/>
       </NP>
       <PREP value="about">
         <SELRESTRS/>
       </PREP>
       <NP value="Topic">
         <SYNRESTRS/>
       </NP>
```

Word Confusion Networks

- Nodes combined to form a linear sequence
- Arcs labeled with words and probabilities
- 1 arc into each node labeled with ε with probability
- Probabilities on the arcs into a node sum to 1.



Parsing Word Confusion Networks

- Modified Gemini parser to handle WCNs
  - Track and combine probabilities
  - Prune phrases with probability beneath a threshold
  - Competing words treated like lexical ambiguity
  - Parser extended to allow ε-moves:
    - For an ε-move between index *i-1* and index *i* with probability p_ε
    - Extend every phrase ending at index *i-1* with probability p_{i-1} to index *i* with probability $p_i = p_{i-1} * p_\varepsilon$
- Parser speed is influenced by
  - Pruning threshold
  - Timeout on the amount of time spent at any index

Evaluation (parser speed)

- Parsed one ICSI meeting (Buw001), 1800 WCNs
  - 31% Word Error Rate
  - Failed to find any major phrases for 177 WCNs
  - WCNs from SRI/ICSI recognizer

Avg. Parse Time	6.5 seconds
Avg. number of nodes	15
Avg. number of arcs	157
Avg. number of phrases	12.7
Avg. phrase length	3.7
Avg. number of edges	478

Evaluation (parser quality)

- Annotaters selected 145 phrases from Buw001 that contribute information relevant to action items
- Judged parser results for each phrase:
  - Identified by parser, with essentially correct semantics
  - Partially identified by parser, but with significant errors or omissions
  - Not identified by parser

Correct	Partial	Missed
35	61	49

Partially Correct Example

- An exampled judged partially correct:
  - Target phrase:

    People are supposed to send me URLs
  - Identified phrase:

    People are supposed to send me elves

- Clearly wrong, but got a lot of the semantics right
- Potentially still useful in the CALO environment

Continuing and Future Work

- Inconsistent use of contracted forms in WCNs
    - Costing us most negations
- Combine lexicon with TRIPS lexicon (U. Rochester)
- N-N modification
    - POS Tag ICSI data to learn common compounds
- Combine WCN probabilities with
    - POS probabilities
    - Parse probabilities
- Evaluate using parser to reduce Word Error Rate