Motivation

• A semantic lexicon assigns semantic categories to words.
  - politician -> human
  - truck -> vehicle
  - grenade -> weapon

• Domain-specific vocabulary is often not found in general purpose resources, such as WordNet.

• Automatic methods could be used to enhance these resources or create domain-specific lexicons.

Syntactic Heuristics for Learning Semantic Labels

- Conjunctions: lions and tigers and bears
- Lists: lions, tigers, bears
- Appositives: the horse, a stallion
- Predicate Nominals: the wolf is a mammal
- Compound nouns: tuna fish, Honda Sedan

[Riloff & Shepherd 97; Roark & Charniak 98; Phillips & Riloff 02; etc.]

- Hyponym patterns: dogs such as beagles and boxers
  - dogs, including beagles and boxers

[ Hearst 92: KnowItAll (U. Washington), Kozareva et al. 2008; etc.]

Bootstrapping Semantic Lexicons

Unannotated Texts

Ex: dog, cat, lion, lizard, snake

N best words

Ex: terrier, poodle, tiger, frog, iguana

Co-occurrence Statistics

prospective category words

Extraction Patterns

- Represent syntactic context that often reveals the semantic class of a word.

- AutoSlog: each pattern extracts an NP from one of 3 syntactic positions: subject, direct object, pp obj.

Some patterns to extract locations:

<subject> was inhabited
  - the locality was inhabited...
patrolling <direct object>
  - ...patrolling Zacamil neighborhood
lives in <pp obj>
  - ...lives in Argentina
Key Ideas behind Basilisk

- Collective evidence over extraction patterns.
- Learning multiple categories simultaneously.

**BASILISK** = _Bootstrapping Approach to Semantic Lexicon Induction using Semantic Knowledge_

The Pattern Pool

Every extraction pattern is scored and the best patterns are put into a _Pattern Pool_.

The scoring function is:

$$ R_{logF \text{ (pattern}_i) = \frac{F_i}{N_i} \times \log_2 (F_i) $$

where:
- $F_i$ is the number of category members extracted by pattern$_i$
- $N_i$ is the total number of nouns extracted by pattern$_i$

The Candidate Word Pool

- All Pattern Pool extractions (head nouns) are added to the Candidate Word Pool.
- Initially, we used a Pattern Pool of size 20, but the pool became stagnant over time.

**Solution:** increase the pool size by 1 after each iteration to infuse the pool with new candidates.
Selecting Words for the Lexicon

Score: the average number of category members extracted by each pattern that extracted the candidate word.

\[
\text{score (word}_i) = \frac{\sum_{j=1}^{N_i} F_j}{N_i}
\]

\[
\text{AvgLog (word}_i) = \frac{\sum_{j=1}^{N_i} \log_2 (F_j + 1)}{N_i}
\]

where:

- \(F_j\) is the number of category members extracted by pattern \(j\)
- \(N_i\) is the total number of patterns that extract word \(i\)

Experimental Design

- Used the MUC-4 corpus: 1700 texts related to terrorism.
- Experiments on 6 semantic categories: 
  
  building, event, human, location, time, weapon.
- 10 seed words for each category.
- 1000 words automatically generated for each category.
- Basilisk compared with our previous algorithm (meta-bootstrapping), adapted to learn nouns instead of NPs.

Baseline Results

Head Nouns (8460 words)

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>building</td>
<td>188</td>
<td>2.2%</td>
</tr>
<tr>
<td>event</td>
<td>501</td>
<td>5.9%</td>
</tr>
<tr>
<td>human</td>
<td>1856</td>
<td>21.9%</td>
</tr>
<tr>
<td>location</td>
<td>1018</td>
<td>12.0%</td>
</tr>
<tr>
<td>time</td>
<td>112</td>
<td>1.3%</td>
</tr>
<tr>
<td>weapon</td>
<td>147</td>
<td>1.7%</td>
</tr>
<tr>
<td>(other)</td>
<td>4638</td>
<td>54.8%</td>
</tr>
</tbody>
</table>

Seed Words

We used the 10 most frequent words for each category.

<table>
<thead>
<tr>
<th>Category</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building</td>
<td>embassy, office, headquarters, church, offices, house, home, residence, hospital, airport</td>
</tr>
<tr>
<td>Event</td>
<td>attack, actions, war, meeting, elections, murder, attacks, action, struggle, agreement</td>
</tr>
<tr>
<td>Human</td>
<td>people, guerrillas, members, troops, Cristiani, rebels, president, terrorists, soldiers, leaders</td>
</tr>
<tr>
<td>Location</td>
<td>country, El Salvador, Salvador, United States, area, Colombia, city, countries, department, Nicaragua</td>
</tr>
<tr>
<td>Time</td>
<td>time, years, days, November, hours, night, morning, week, year, day</td>
</tr>
<tr>
<td>Weapon</td>
<td>weapons, bomb, bombs, explosives, arms, missles, dynamite, rifles, materiel, bullets</td>
</tr>
</tbody>
</table>
Learning Multiple Categories Simultaneously

- We hypothesized that confusion errors can be reduced by learning multiple semantic categories simultaneously.
- “One Sense per Domain” assumption.
- Knowledge about competing categories can constrain and steer the bootstrapping process.

Semantic Learning Case Study

- Input to Basilisk: 10 common disease names
- Of the top 200 words hypothesized to be diseases: 89 were already in the UMLS metathesaurus (32,000 names of diseases and organisms), but 111 were not! Including:
  
<table>
<thead>
<tr>
<th>Disease Name</th>
<th>Disease Name</th>
<th>Disease Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>adenomatosis</td>
<td>flu</td>
<td>h5n1</td>
</tr>
<tr>
<td>tularaemia</td>
<td>kawasaki</td>
<td>h7n3</td>
</tr>
<tr>
<td>tularamia</td>
<td>mad-cow-disease</td>
<td>ev71</td>
</tr>
<tr>
<td>diarrhoea</td>
<td>smut</td>
<td>yf</td>
</tr>
<tr>
<td>diphtheriae</td>
<td>pertussis</td>
<td>jyf</td>
</tr>
<tr>
<td>enterovirus-71</td>
<td>pleuro-pneumonia</td>
<td>nvcjd</td>
</tr>
<tr>
<td>fibropapillomas</td>
<td>polioencephalomyelitis</td>
<td>pepmv</td>
</tr>
<tr>
<td>gastroenteritis</td>
<td>poliovirus</td>
<td>wsmv</td>
</tr>
</tbody>
</table>

Learning Multiple Categories Simultaneously
Bootstrapping Multiple Categories

Simple Conflict Resolution

• A word cannot be assigned to category X if it has already been assigned to category Y.

• If a word is hypothesized for both category X and category Y at the same time, choose the category that receives the highest score.

Building

Human

Event

Weapon

Time

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1

Correct Lexicon Entries

Total Lexicon Entries

BA-M

BA-1
A Smarter Scoring Function

A more proactive approach: incorporate knowledge about other categories directly into the scoring function.

New scoring function:

$$\text{diff} (w_i, c_a) = \text{AvgLog} (w_i, c_a) - \max (\text{AvgLog}(w_i, c_b))$$

Conclusions

- Using collective evidence from a set of extraction patterns improves the accuracy of semantic lexicon induction.
- Learning multiple semantic categories at the same time can constrain bootstrapping and improve performance.
- Basilisk is much faster than meta-bootstrapping since Basilisk uses only a single bootstrapping loop.
- Manual review is still necessary to use the learned dictionaries.
- Performance for some categories is beginning to approach levels for which manual review may not be necessary.