Motivation

- A semantic lexicon assigns semantic categories to words.
- 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
- the horse, a stallion

Lists

- lions, tigers, bears
- the wolf is a mammal

Appositives

- tuna fish
- Honda Sedan

Predicate Nominals

- Compound nouns

Hyponym patterns

- dogs such as beagles and boxers
- dogs, including beagles and boxers

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

Bootstrapping Semantic Lexicons

Unannotated Texts

- Ex: dog, cat, lion, lizard, snake

Co-occurrence Statistics

- N best words
- Ex: terrier, poodle, tiger, frog, iguana
Mutual Bootstrapping [Riloff & Jones 99]

Unannotated Texts

Ex: dog, cat, lion, lizard, snake

Best Extraction Pattern

Ex: Rottweiler, terrier, cougar

Extractions (Nouns)

Examples of Learned Patterns

<table>
<thead>
<tr>
<th>Location Patterns (Web)</th>
<th>Location Patterns (Terrorism)</th>
</tr>
</thead>
<tbody>
<tr>
<td>offices in &lt;np&gt;</td>
<td>living in &lt;np&gt;</td>
</tr>
<tr>
<td>facilities in &lt;np&gt;</td>
<td>traveled in &lt;np&gt;</td>
</tr>
<tr>
<td>operations in &lt;np&gt;</td>
<td>become in &lt;np&gt;</td>
</tr>
<tr>
<td>loans in &lt;np&gt;</td>
<td>sought in &lt;np&gt;</td>
</tr>
<tr>
<td>operates in &lt;np&gt;</td>
<td>presidents in &lt;np&gt;</td>
</tr>
<tr>
<td>locations in &lt;np&gt;</td>
<td>parts of &lt;np&gt;</td>
</tr>
<tr>
<td>producer in &lt;np&gt;</td>
<td>to enter &lt;np&gt;</td>
</tr>
<tr>
<td>states of &lt;np&gt;</td>
<td>condemned in &lt;np&gt;</td>
</tr>
<tr>
<td>seminars in &lt;np&gt;</td>
<td>relations between &lt;np&gt;</td>
</tr>
<tr>
<td>activities in &lt;np&gt;</td>
<td>ministers of &lt;np&gt;</td>
</tr>
<tr>
<td>consulting in &lt;np&gt;</td>
<td>part in &lt;np&gt;</td>
</tr>
<tr>
<td>countries of &lt;np&gt;</td>
<td>taken in &lt;np&gt;</td>
</tr>
</tbody>
</table>

Mutual Bootstrapping Example

SEEDS: Nicaragua, city, region, town

Best pattern: headquartered in <NP>
Extractions: Nicaragua, city, Chapare region, San Miguel

Best pattern: downed in <NP>
Extractions: Nicaragua, city, Usulutan region, San Miguel, area, Soyapango

Best pattern: to occupy <NP>
Extractions: Nicaragua, town, this northern area, small country, San Sebastian neighborhood, private property
BASILISK = Bootstrapping Approach to Semantic Lexicon Induction using Semantic Knowledge

Key Ideas behind Basilisk

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

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.

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

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

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

<table>
<thead>
<tr>
<th>Head Nouns (8460 words)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>building</td>
<td>188 (2.2%)</td>
</tr>
<tr>
<td>event</td>
<td>501 (5.9%)</td>
</tr>
<tr>
<td>human</td>
<td>1856 (21.9%)</td>
</tr>
<tr>
<td>location</td>
<td>1018 (12.0%)</td>
</tr>
<tr>
<td>time</td>
<td>112 (1.3%)</td>
</tr>
<tr>
<td>weapon</td>
<td>147 (1.7%)</td>
</tr>
<tr>
<td>(other)</td>
<td>4638 (54.8%)</td>
</tr>
</tbody>
</table>

Seed Words

We used the 10 most frequent words for each category.

- **Building**: embassy, office, headquarters, church, offices, house, home, residence, hospital, airport
- **Event**: attack, actions, war, meeting, elections, murder, attacks, action, struggle, agreement
- **Human**: people, guerrillas, members, troops, Cristiani, rebels, president, terrorists, soldiers, leaders
- **Location**: country, El Salvador, Salvador, United States, area, Colombia, city, countries, department, Nicaragua
- **Time**: time, years, days, November, hours, night, morning, week, year, day
- **Weapon**: weapons, bomb, bombs, explosives, arms, missiles, dynamite, rifles, materiel, bullets
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:
  adenomatosis  flu  h5n1
tularaemia  kawasaki  h7n3
tularamia  mad-cow-disease  ev71
diarrhoea  smut  yf
diphtheriae  pertussis  jyf
enterovirus-71  pleuro-pneumonia  nvcjd
fibropapillomas  polioencephalomyelitis  pepmv
gastroenteritis  poliovirus  wsmv

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.
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.
A Smarter Scoring Function

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

New scoring function:

\[ \text{diff} (w_i, c_a) = \text{AvgLog} (w_i, c_a) - \max_{b \neq a} (\text{AvgLog}(w_i, c_b)) \]

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.