Automatic Lexical Acquisition

Paola Merlo
University of Geneva

Why automatic lexical acquisition?

- TEXT MINING
- TEXT CLASSIFICATION
- QUESTION ANSWERING
- DIALOGUE MANAGEMENT
- LARGE DETAILED LEXICONS
- TEXT SUMMARISATION
- INFORMATION EXTRACTION
- TEMPLATE FILLING

Verb classification

- Verbs are the primary source of relational information in a sentence
  
  Jane hit the ball
  
  NP Agent Theme

- Classification as indirect learning of the lexicon for
  
  - easy organisation: verbs can be organised around shared syntactic and semantic properties
  - consistent extension: associating a verb with a class allows it to inherit detailed linguistic information

Example of verb classification

- English verb classes according to Levin
  
  approximately 200 classes for 3000 verbs

- For example

  Manner of Motion: race, jump, skip, moosey
  Sound Emission: buzz, ring, crack
  Change of State: burn, melt, pour
  Creation/Transformation: build, carve
  Psychological state: admire, love, hate, despise
**Verb alternations**

How does one reach such a classification?

Hypothesis: verbs with a similar semantics express their arguments in a similar way. They exhibit alternations.

Example:

- if a verb can be transitive: melt butter jump horse
- and it can be intransitive: butter melts horse jumps
- and it can have an adjectival form: melted butter *jumped horse
- then it is a verb of change of state

**Related Work**

Syntactic information -- subcategorization frames
- machine readable dictionary (Dorr 97)
- examples of usage in a corpus (Brent 93, Briscoe and Carroll 97, McCarthy and Korhonen 98, Korhonen 2000, 2002, Lapata 99, Manning 93)

Semantic information
- selectional restrictions (Resnik 96)
- verbal aspect (Siegel and McKeown 2001);

**Our Proposal** (Merlo and Stevenson 2001)

- Verbs which share semantic properties also share syntactic properties
- There is a regular mapping from meaning components to syntactic usage (Levin 93, Pinker 89)
- Can reason in reverse direction and induce semantic class from syntactic usage

**Methodology**

- Analyse verb classes to determine discriminating thematic properties
- Develop indicators (indicator random variables) that approximate thematic properties and that can be counted in a corpus
- Collect relative frequencies to generate a statistical summary of the thematic behaviour of each verb
- Apply machine learning algorithm (e.g. decision tree induction) to produce a classifier
**English Verb Classes**

Three classes of optionally intransitive verbs

<table>
<thead>
<tr>
<th>Manner of Motion</th>
<th>The rider raced the horse past the barn (Causal) Agent</th>
</tr>
</thead>
<tbody>
<tr>
<td>The horse raced past the barn Agent</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change of State</th>
<th>The cook melted the butter (Causal) Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>The butter melted Theme</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Creation/Transformation</th>
<th>The contractors built the house Agent Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>The contractors built all summer Agent</td>
<td></td>
</tr>
</tbody>
</table>

**Summary of Thematic Assignments**

<table>
<thead>
<tr>
<th>Classes</th>
<th>Transitive</th>
<th>Intransitive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Subject</td>
<td>Object</td>
</tr>
<tr>
<td>Manner of Motion (race)</td>
<td>(Causal) Agent</td>
<td>Agent</td>
</tr>
<tr>
<td>Change of State (melt)</td>
<td>(Causal) Agent</td>
<td>Theme</td>
</tr>
<tr>
<td>Create/Transform (build)</td>
<td>Agent</td>
<td>Theme</td>
</tr>
</tbody>
</table>

**MAIN IDEA**

Underlying thematic differences among the verb classes will surface as detectable differences in the usage of surface indicators.

**Features for Automatic Classification: Example**

<table>
<thead>
<tr>
<th>Classes</th>
<th>Transitive</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoM</td>
<td>(Causal) Agent</td>
<td>The jockey raced the horse Agent</td>
</tr>
<tr>
<td>CoS</td>
<td>(Causal) Agent Theme</td>
<td>The cook melted the butter</td>
</tr>
<tr>
<td>C/T</td>
<td>Agent Theme</td>
<td>The workers built the house</td>
</tr>
</tbody>
</table>

**Relationship between Frequency and Transitivity**

- **Transitivity by causation: MoM, CoS**
  - Greater complexity, two events

- **Agentive object : MoM** (transitive unergative)
  - Infrequent in English: only MoM and SE
  - Infrequent typologically (* Italian, French, German, Portuguese, Gungbe and Czech. Vietnamese only comitative)
  - Difficult to process (Bever 1970, Stevenson Merlo 97, Filip et al. CUNY 98)

  - **Expected frequency of transitive use** MoM < CoS < C/T
### Features for Automatic Classification (2/3)

<table>
<thead>
<tr>
<th>Classes</th>
<th>Object of Transitive</th>
<th>Object of Intransitive</th>
<th>Subject of Transitive</th>
<th>Subject of Intransitive</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoM</td>
<td>Agent</td>
<td>Agent</td>
<td>The jockey raced the horse</td>
<td>The horse raced</td>
<td></td>
</tr>
<tr>
<td>CoS</td>
<td>Theme</td>
<td>Theme</td>
<td>The cook melted the butter</td>
<td>The butter melted</td>
<td></td>
</tr>
<tr>
<td>C/T</td>
<td>Theme</td>
<td>Agent</td>
<td>no alternation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Feature**: Causativity.
Amount of overlap between subject of intransitive and object of transitive.

### Features for Automatic Classification (3/3)

<table>
<thead>
<tr>
<th>Classes</th>
<th>Subject of Transitive</th>
<th>Subject of Intransitive</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoM</td>
<td>Causer</td>
<td>Agent</td>
<td>The jockey raced the horse</td>
</tr>
<tr>
<td>CoS</td>
<td>Causer</td>
<td>Theme</td>
<td>The cook melted the butter</td>
</tr>
<tr>
<td>C/T</td>
<td>Agent</td>
<td>Agent</td>
<td>The workers built</td>
</tr>
</tbody>
</table>

**Feature**: Animacy
Themes are more likely to be inanimate.

### Summary of Expectations of Features

- **Transitivity**: MoM < CoS < C/T
- **Causativity**: CoS > {MoM, C/T}
- **Animacy**: CoS < {MoM, C/T}

### Indicator Random Variables for Transitivity

**TRANS**:
- 1 if verb is used transitively
- 0 if verb is used intransitively

**PASS**:
- 1 if verb is passive
- 0 if verb is active

**VBN**:
- 1 if verb is past participle
- 0 if verb is not past participle
**Indicator Random Variables for Animacy**

\[
\text{ANIM}_v; \begin{cases} 
1 & \text{if subject of verb is animate} \\
0 & \text{if subject of verb is inanimate}
\end{cases}
\]

Animacy is approximated by personal pronouns

**Indicator Random Variables for Causativity**

Let a sample space of pairs of transitive objects and intransitive subjects of the verb be given. We define the CAUS indicator random variable for the verb as follow:

\[
\text{CAUS}_v; \begin{cases} 
1 & \text{if subject = object} \\
0 & \text{otherwise}
\end{cases}
\]

**Probabilities**

Probabilities of random variables are estimated by simple relative frequencies

Example

\[
P(\text{TRANS}) = \frac{C(v,o)}{C(v,o) + C(v,0)}
\]

Occurrences of verb followed by object over total occurrences of verb, followed by object or not

Vector template: [ verb,TRANS,PASS,VBN,CAUS,ANIM,class]

Example: [ open, .69, .09, .21, .16, .36, CoS ]

**Data Collection -- Method (1/2)**

TRANS

Verb token immediately followed by potential object counted as transitive else intransitive.

Potential object = Closest nominal group after verb token.

PASS

Main verb (VBD) = active.

Token with tag VBN counted as active, if closest preceding auxiliary was have, counted as passive if closest preceding auxiliary was be.

VBN

POS label according to the tagged corpus.
**Data Collection -- Method (2/2)**

**CAUS**

- Extract multiset of subjects and multiset of objects for each verb.
- Calculate overlap of two multisets.
- Take ratio between cardinality of the overlap multiset, and the sum of the cardinality of the subject and object multisets.

**ANIM**

- Ratio of occurrences of pronoun subjects to all subjects for each verb.

**Statistical Analysis of the Data**

Mean relative frequencies

<table>
<thead>
<tr>
<th></th>
<th>TRANS</th>
<th>PASS</th>
<th>VBN</th>
<th>CAUS</th>
<th>ANIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoM</td>
<td>.23</td>
<td>.07</td>
<td>.12</td>
<td>.00</td>
<td>.25</td>
</tr>
<tr>
<td>CoS</td>
<td>.40</td>
<td>.33</td>
<td>.27</td>
<td>.12</td>
<td>.07</td>
</tr>
<tr>
<td>ObD</td>
<td>.62</td>
<td>.31</td>
<td>.26</td>
<td>.04</td>
<td>.15</td>
</tr>
</tbody>
</table>

All statistically significant at p< .01, except the difference between CoS and ObD for PASS and VBN

**English Supervised Experiments**

**Materials**

- 59 verbs (20 MoM, 19 CoS, 20 C/T)
- 65 million tagged words (29 million parsed) (WSJ and Brown corpus)

**Method**

- Learner: C5.0 (decision tree induction algorithm)
- Training/Testing: 10-fold cross-validation repeated 50 times

**Results**

- **Overall results**: accuracy 69.8% (baseline 33.9, expert upper bound 86.5%)
  (recent replication on chunked BNC accuracy 82.4%)

- 54% reduction in error rate on previously unseen verbs

- **Effectiveness of features**
  All features, except PASS, are useful in classification

- **Class by class accuracy**
  MoM verbs are most accurately classified

- **Analysis of errors**
  Hypothesized relation between features and thematic assignments is confirmed
Results

• Overall results: accuracy 69.8% (baseline 34%, expert upper bound 86.5%) (recent replication on chunked BNC accuracy 82.4%)

• Effectiveness of features:
  All features, except PASS, are useful in classification

<table>
<thead>
<tr>
<th>FEATURES</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 TRANS PASS VBN CAUS ANIM</td>
<td>69.8</td>
</tr>
<tr>
<td>2 TRANS PASS VBN CAUS ANIM</td>
<td>69.8</td>
</tr>
<tr>
<td>3 TRANS PASS VBN CAUS ANIM</td>
<td>67.3</td>
</tr>
<tr>
<td>4 TRANS PASS VBN CAUS ANIM</td>
<td>66.5</td>
</tr>
<tr>
<td>5 TRANS PASS VBN CAUS ANIM</td>
<td>63.2</td>
</tr>
<tr>
<td>6 TRANS PASS VBN CAUS ANIM</td>
<td>61.6</td>
</tr>
</tbody>
</table>

Class by Class Results

<table>
<thead>
<tr>
<th>FEATURES</th>
<th>UNUSED</th>
<th>MoM</th>
<th>CoS</th>
<th>C/T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 TRANS PASS VBN CAUS ANIM</td>
<td>73.9</td>
<td>68.6</td>
<td>64.9</td>
<td></td>
</tr>
<tr>
<td>2 TRANS VBN CAUS ANIM</td>
<td>PASS</td>
<td>76.2</td>
<td>75.7</td>
<td>61.6</td>
</tr>
<tr>
<td>3 TRANS PASS VBN ANIM</td>
<td>CAUS</td>
<td>65.1</td>
<td>60.0</td>
<td>62.8</td>
</tr>
<tr>
<td>4 TRANS PASS CAUS ANIM</td>
<td>VBN</td>
<td>66.7</td>
<td>65.0</td>
<td>51.3</td>
</tr>
<tr>
<td>5 TRANS PASS VBN CAUS</td>
<td>ANIM</td>
<td>72.7</td>
<td>47.0</td>
<td>60</td>
</tr>
<tr>
<td>6 PASS VBN CAUS ANIM</td>
<td>TRANS</td>
<td>78.1</td>
<td>51.5</td>
<td>61.9</td>
</tr>
</tbody>
</table>

Analysis of Errors

• TRANS sharpens 3 way distinction
• ANIM particularly helpful in discriminating CoS
• VBN (past participle) primarily discriminates C/T

Conclusion

• Hypothesis confirmed
  corpus-based indicators reflect underlying semantic properties of verbs

• Method has high performance
**Generalising to a new class**

- **New Class**  Psychological State Verbs
- **New thematic roles**  Experiencer Stimulus

**Example**

<table>
<thead>
<tr>
<th>Experiencer</th>
<th>Stimulus</th>
</tr>
</thead>
<tbody>
<tr>
<td>The rich</td>
<td>love</td>
</tr>
<tr>
<td>The rich</td>
<td>love too</td>
</tr>
</tbody>
</table>

**Indicators**: TRANS, CAUS, ANIM

PROG use of the progressive (stative/non stative)

carefully indicator of volitionality (agent vs experiencer)

**Classes**: MoM, CoS, C/T, Psy

**Results and Discussion**

**Results**

- 75.6% accuracy (baseline 57.4%)
- 43% reduction in error rate

TRANS, ANIM, PROG, Carefully best features

Relationship between indicators and thematic properties holds across classes

Some specific indicators carry across thematic roles

**Discovery**

We do not need to investigate new indicators for each new class (73.5% accuracy with only old indicators)

Conjecture: Indicators are partially correlated with thematic roles and they capture commonalities across roles

**Relevance for acquisition of verb meaning (1/3)**

(Stevenson and Merlo CUNY 2001)

**Syntactic Bootstrapping**

The acquisition of a verb's meaning is constrained by the verb's linguistic contexts -- its subcategorisation frames (Gleitman 1990) and its argument structure (Gillette et al. 1999).

**Question**

How does the learner induce subcategorisation and argument structure information?

**Relevance for acquisition of verb meaning (2/3)**

**Our proposal**

Argument structure distinctions can be learnt from simple syntactic information

- frequencies of subcategorization frames
- alternations in the realisation of arguments (requires correspondences across subcategorization frames)
- other alignments between syntax and semantics: animacy

**Results of unsupervised experiments** (hierarchical clustering)

Indicators distinguish classes at 63% accuracy
On-going and Future Research

• **NLP** - more languages (Italian, German, Chinese)
- more learning features (aspect)
- automatic distinction of arguments from adjuncts
  (Merlo EACL'03)

• **Machine Learning**
  - generic feature space (Joanis)
  - multi-lingual classification using co-training
  - unsupervised clustering of Spanish verbs (Esteve Ferrer)

• **Applications** - enriching document representations for summarisation

THANK YOU