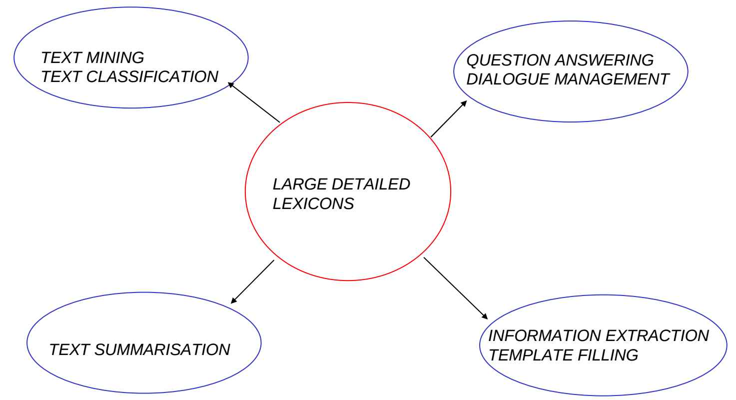
Automatic Lexical Acquisition

Paola Merlo
University of Geneva

Why automatic lexical acquisition?



Verb classification

- Verbs are the primary source of relational information in a sentence

Jane	*hit*	*the ball*
NP		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);
- lexical semantic classes (Lapata and Brew 99, Schulte im Walde 2000, McCarthy and Korhonen 2000, Merlo and Stevenson 2001, Gildea and Jurafsky 2002)

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

Learn verb classes based on thematic relations using only corpus-based statistics

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

Manner of Motion	**The rider** *(Causal) Agent* **raced** **the horse** *Agent* **past the barn**
	The horse *Agent* **raced past the barn**
Change of State	**The cook** *(Causal) Agent* **melted** **the butter** *Theme*
	The butter *Theme* **melted**
Creation/Transformation	**The contractors** *Agent* **built** **the house** *Theme*
	The contractors *Agent* **built all summer**

Summary of Thematic Assignments

Classes	Transitive		Intransitive
	Subject	Object	Subject
Manner of Motion *(race)*	(Causal) Agent	Agent	Agent
Change of State *(melt)*	(Causal) Agent	Theme	Theme
Create/Transform *(build)*	Agent	Theme	Agent

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

Classes	Transitive		Example
	Subject	Object	
MoM	(Causal) Agent	Agent	The jockey raced the horse
CoS	(Causal) Agent	Theme	The cook melted the butter
C/T	Agent	Theme	The workers built the house

Feature Transitivity (usage in the transitive construction)

- **Transitivity by causation is more complex**
- **Agent object is (typologically) rare**

- **Expected order of Transitivity: MoM < CoS < C/T**

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)

Classes	Object of Transitive	Subject of Intransitive	Example
MoM	Agent	Agent	The jockey raced the horse The horse raced
CoS	Theme	Theme	The cook melted the butter The butter melted
C/T	Theme	Agent	no alternation

Feature Causativity.

Amount of overlap between subject of intransitive and object of transitive

Features for Automatic Classification (3/3)

Classes	Subject of Transitive	Subject of Intransitive	Example
MoM	Causer	Agent	The jockey raced the horse The horse raced
CoS	Causer	Theme	The cook melted the butter The butter melted
C/T	Agent	Agent	The workers built The workers built the house

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

$$\text{TRANS}_v: \begin{cases} 1 & \text{if verb is used transitively} \\ 0 & \text{if verb is used intransitively} \end{cases}$$

$$\text{PASS}_v: \begin{cases} 1 & \text{if verb is passive} \\ 0 & \text{if verb is active} \end{cases}$$

$$\text{VBN}_v: \begin{cases} 1 & \text{if verb is past participle} \\ 0 & \text{if verb is not past participle} \end{cases}$$

Indicator Random Variables for Animacy

ANIM$_v$: $\begin{cases} \textbf{1 if subject of verb is animate} \\ \textbf{0 if subject of verb is inanimate} \end{cases}$

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

CAUS$_v$: $\begin{cases} \textbf{1 if subject = object} \\ \textbf{0 otherwise} \end{cases}$

Probabilities

Probabilities of random variables are estimated by simple relative frequencies

Example

$$P(TRANS_v) \approx \frac{C(v,o)}{C(v,o)+C(v,\bar{o})}$$

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

	TRANS	PASS	VBN	CAUS	ANIM
MoM	.23	.07	.12	.00	.25
CoS	.40	.33	.27	.12	.07
ObD	.62	.31	.26	.04	.15

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% (baseline34%,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

	FEATURES	Accuracy %
1	TRANS PASS VBN CAUS ANIM	69.8
2	TRANS ~~PASS~~ VBN CAUS ANIM	69.8
3	TRANS PASS VBN ~~CAUS~~ ANIM	67.3
4	TRANS PASS ~~VBN~~ CAUS ANIM	66.5
5	TRANS PASS VBN CAUS ~~ANIM~~	63.2
6	~~TRANS~~ PASS VBN CAUS ANIM	61.6

Class by Class Results

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

- MoM are the best identified

Analysis of Errors

ALL FEATURES				
		Assigned Class		
		MoM	CoS	C/T
Actual Class	MoM		1	2
	CoS	4		3
	C/T	5	3	

Without Animacy				
		Assigned Class		
		MoM	CoS	C/T
Actual Class	MoM		2	2
	CoS	5		6
	C/T	3	5	

- 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

The rich love money
Experiencer *Stimulus*

The rich love too
Experiencer

•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 linguistics 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)

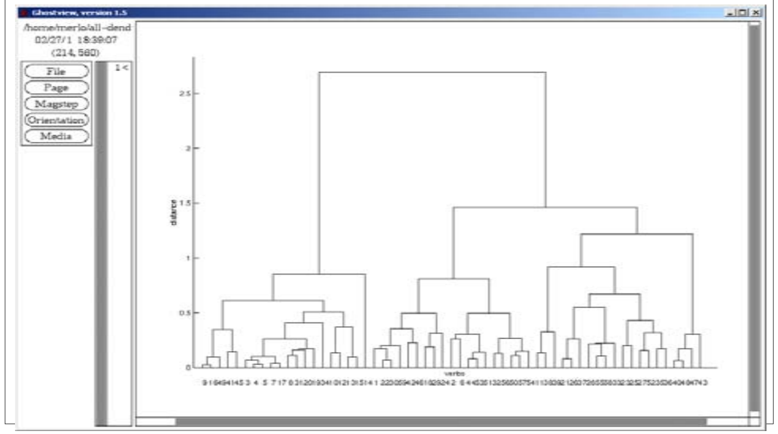
•**Our proposal** Argument structure distinctions can be learnt from simple syntactic information

- 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**

Unsupervised acquisition of verb meaning (3/3)



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