Final Term Review

Semantics
Information Extraction

Semantics

- Concepts
- Word Similarities based on thesauri
- Word Vectors (Sparse, Dense)
- Semantic Role Labeling

Concepts

- Word Meanings
  - Homonymy (Bank: financial institution, river bank)
  - Polysemy (Bank: financial institution, bank building)
  - Metonymy (Bank or School: organization, building)
- Word Relations
  - Synonyms (big / large)
  - Antonyms (big / small)
  - Hyponym (car is a hyponym of vehical)
  - Hypernym (vehical is a hyponym of car)
  - Instance (College Station is a town)

Word Similarity based on Thesauri

- Path based, 1/pathlen
- Information Content, IC (LCS (c1, c2)), -logP(LCS (c1, c2)) (Resnik)
- Improved Information Content, considering both commonality and differences, 2logP(LCS (c1, c2)) / (logP(c1) + logP(c2)) (Dekang Lin)

Word Vectors

- Distributional vectors (sparse)
  - Term-document matrix -> term-term matrix
  - Frequency -> PPMI, log(p(w1,w2)/p(w1)*p(w2))
  - Similarity: Cosine of two word vectors
- Dense vectors
  - Singular Value Decomposition
  - Prediction-based
  - Brown clustering

Semantic Role Labeling

- Semantic roles (thematic roles): the abstract role that arguments of a predicate can take wrt the event represented by the predicate.
- Agent, theme, source, target …
- Propbank, framenet

A simple modern algorithm

function SEMANTICROLELABEL(*words*) **returns** labeled tree

parse ← PARSE(*words*)
for each *predicate* **in** *parse* **do**
  for each *node* **in** *parse* **do**
    featurevector ← EXTRACTFEATURES(*node*, *predicate*, *parse*)
    CLASSIFYNODE(*node*, *featurevector*, *parse*)

Information Extraction

- Semantic Lexicon Induction
- Relation Extraction
- Coreference resolution
- Event Extraction

Semantic Lexicon Induction

- Syntactic Heuristics
- Co-occurrence based Bootstrapping
- Mutual bootstrapping

Syntactic Heuristics for Learning Semantic Labels

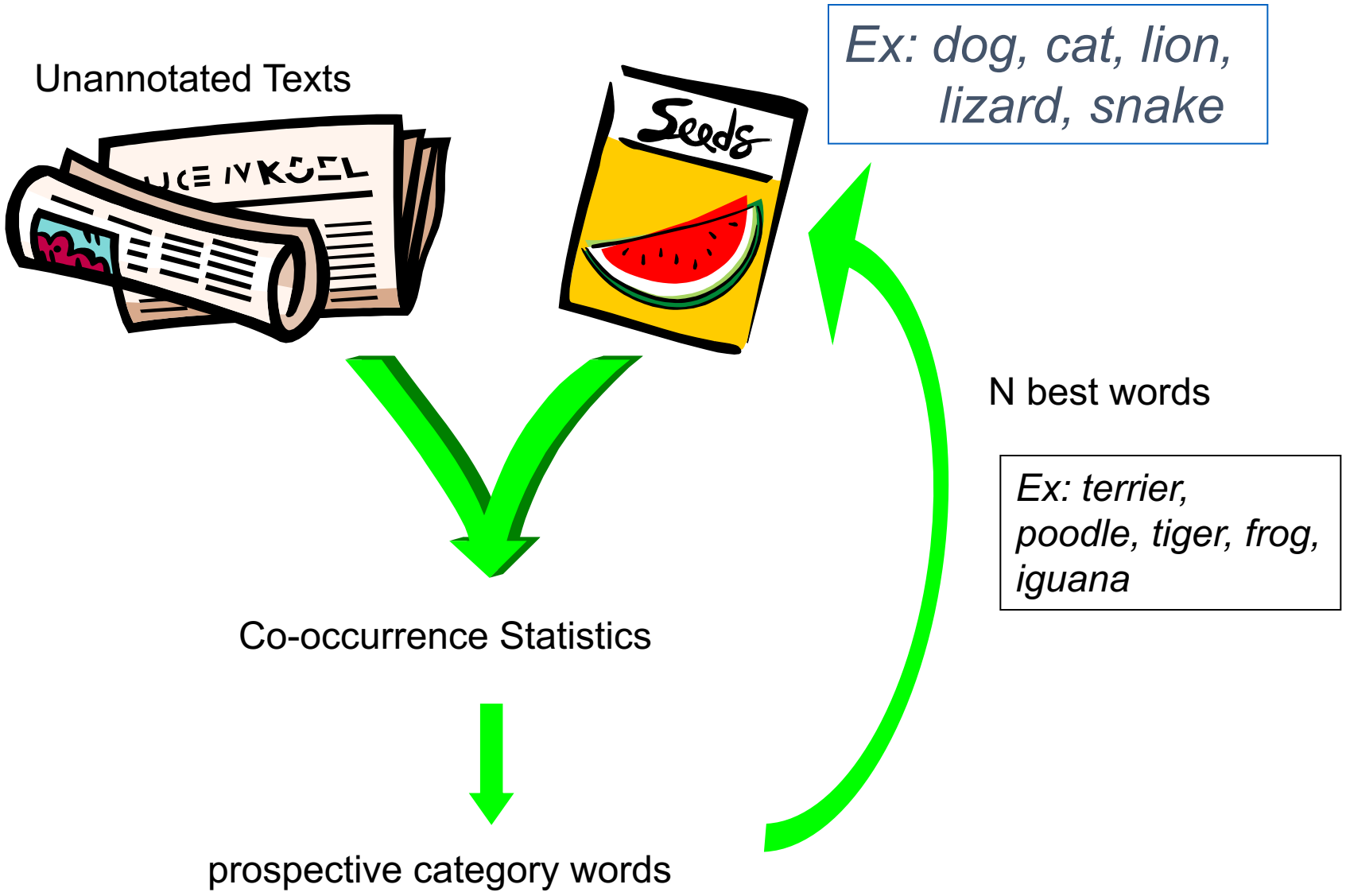
Conjunctions	lions and tigers and bears
Lists	lions, tigers, bears
Appositives	the horse, a stallion
Predicate Nominals	the wolf is a mammal
Compound nouns	tuna fish
	Honda Sedan

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

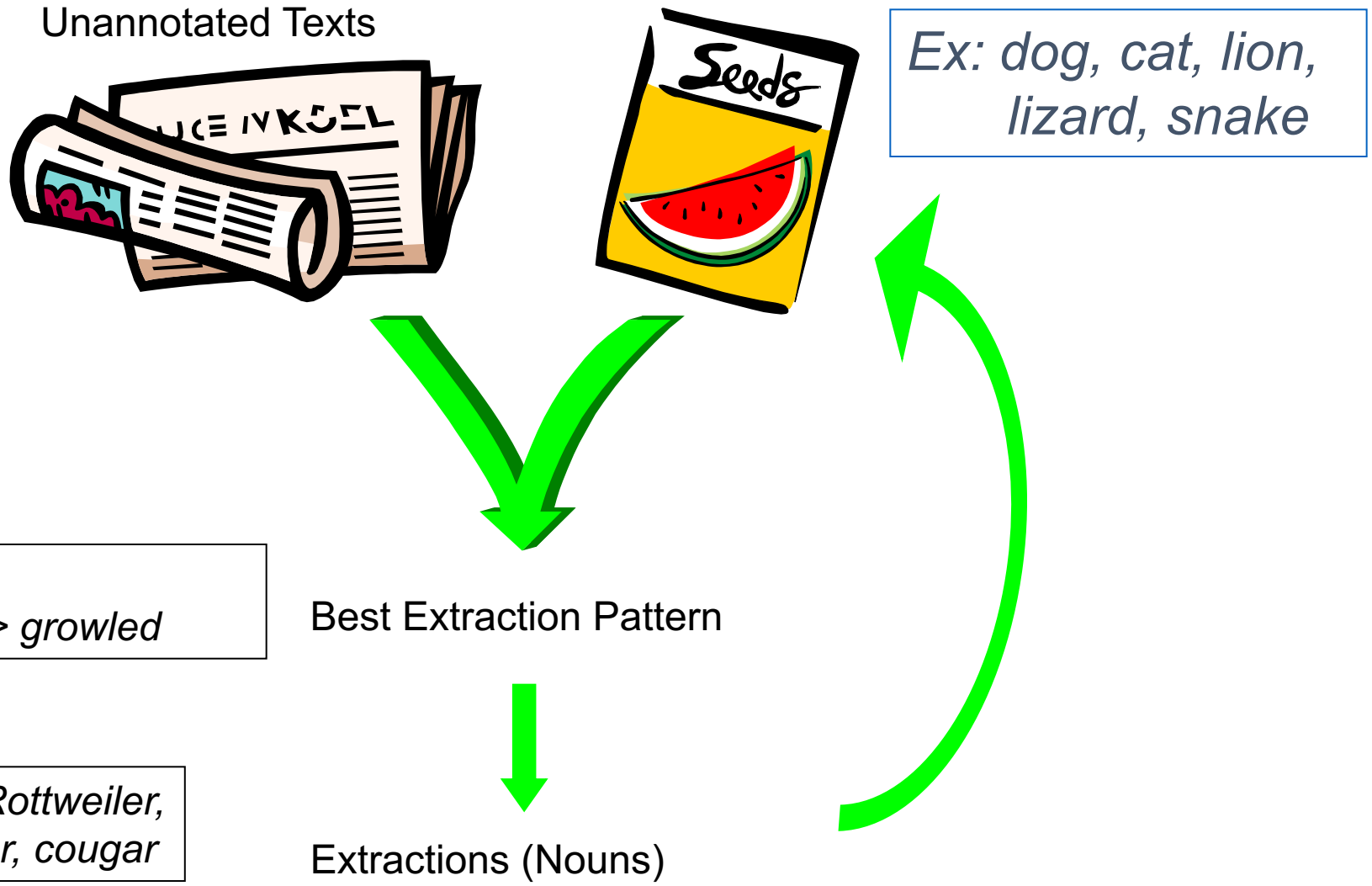
Hyponym patterns	dogs such as beagles and boxers
	dogs, including beagles and boxers

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

Bootstrapping Semantic Lexicons



Mutual Bootstrapping [Riloff & Jones 99]

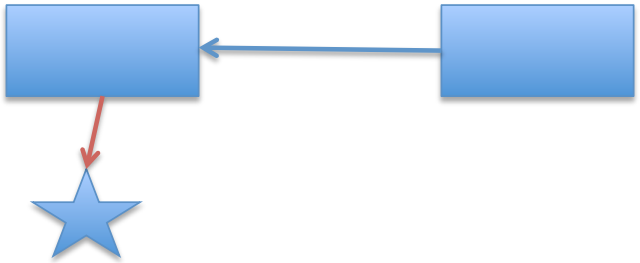


How to build relation extractors

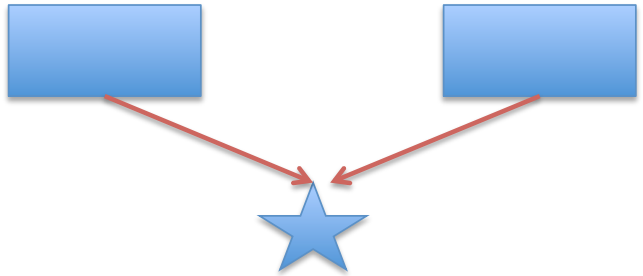
1. Hand-written patterns
2. Supervised machine learning
3. Semi-supervised and unsupervised
   - Bootstrapping (using seeds)
   - Distant supervision
   - Unsupervised learning from the web

Two different things…

- Anaphora
  - Text
  - World

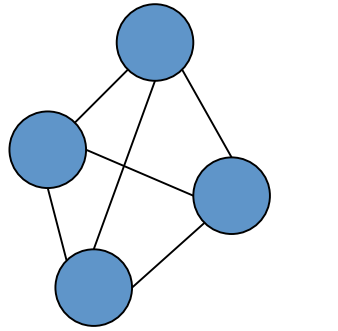


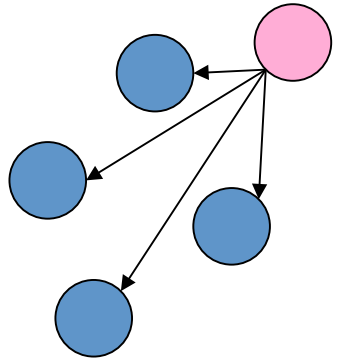
- (Co)Reference
  - Text
  - World



Kinds of Models

- Mention Pair models
  - Treat coreference chains as a collection of pairwise links
  - Make independent pairwise decisions and reconcile them in some way (e.g. clustering or greedy partitioning)
- Mention ranking models
  - Explicitly rank all candidate antecedents for a mention
- Entity-Mention models
  - A cleaner, but less studied, approach
  - Posit single underlying entities
  - Each mention links to a discourse entity [Pasula et al. 03], [Luo et al. 04]





Patterns/Rules vs. Sequence Tagging

Two general approaches to IE:

Pattern-based systems use patterns or rules that are applied to text.

Sequence tagging models classify individual tokens as to whether or not they should be extracted.

AutoSlog-TS [Riloff 96] (Step 1)

Relevant

Irrelevant

[The World Trade Center], [an icon] of [New York City], was horrifically attacked on [an otherwise beautiful day] in [September 2001] by [Al Qaeda].

Shallow Parser

Syntactic Templates →

Extraction Patterns:
<subj> was attacked
icon of <np>
was attacked on <np>
was attacked in <np>
was attacked by <np>

AutoSlog-TS (Step 2)

Relevant

Irrelevant

Extraction Patterns:
<subj> was attacked
icon of <np>
was attacked on <np>
was attacked in <np>
was attacked by <np>

Extraction Patterns	Freq	Prob
<subj> was attacked	100	.90
icon of <np>	5	.20
was attacked on <np>	80	.79
was attacked in <np>	85	.87
was attacked by <np>	95	.95