Building Large-Scale Ontology by Learning from Text

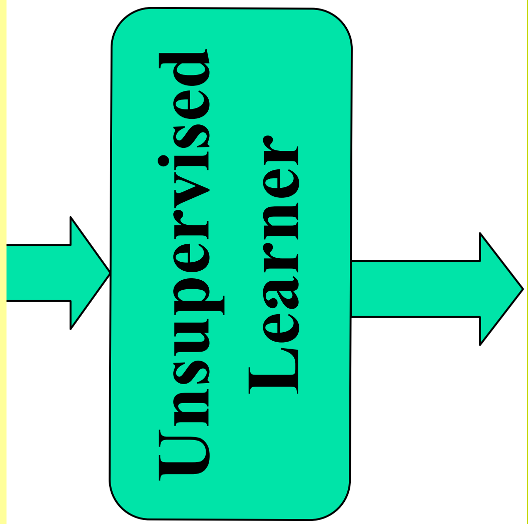
Dekang Lin
Department of Computing Science
University of Alberta
lindek@cs.ualberta.ca

What is an Ontology?

- A set of concepts
- Relations between concepts
- Inference rules among the relations

Unsupervised Learning from Text


```
<DOC>
<DOCNO> AP880212-0006 </DOCNO>
<FILEID>AP-NR-02-12-88 1644EST</FILEID>
<FIRST>r i AM-CagedHens     02-12 0159</FIRST>
<SECOND>AM-Caged Hens,0162</SECOND>
<HEAD>Court Rules Caging Hens Is Not Cruelty</HEAD>
<DATELINE>STROEMMEN, Norway (AP) </DATELINE>
<TEXT>
   A court ruled Friday that an egg
producer who kept his 2,000 hens in small cages was not guilty of
cruelty to animals, as alleged by animal rights activists.
   ``The verdict is a great relief. It would have been too much to
be found guilty of cruelty to my 2,000 hens,'' Karl Wettre was
quoted as saying by the national NTB news agency after his
acquittal.
   The National Society for the Prevention of Cruelty to Animals
claimed that by keeping hens in small cages, Wettre violated
national legislation to allow animals' natural development and
behavior.
   But the court found that Wettre observed Norwegian regulations
stipulating that a hen should have at least 112 square inches of
cage space in which to live.
   NSPCA chairman Toralf Metveit was quoted as saying: ``I'm
disappointed but not surprised.''
   The society was ordered pay $15,600 in court costs.
</TEXT>
</DOC>
<DOC>
<DOCNO> AP880212-0007 </DOCNO>
<FILEID>AP-NR-02-12-88 1518EST</FILEID>
<FIRST>u p AM-Kemp'sStrategy     02-12 0654</FIRST>
<SECOND>AM-Kemp's Strategy,650</SECOND>
<HEAD>Kemp Strategy To Crack Top Three in N.H.
Primary</HEAD>
<HEAD>With AM-Kemp-Robertson Bjt</HEAD>
<BYLINE>By JONATHAN KELLOGG</BYLINE>
<BYLINE>Associated Press Writer</BYLINE>
<DATELINE>NASHUA, N.H. (AP) </DATELINE>
<TEXT>
   Strategists for Jack Kemp's presidential
campaign say George Bush's poor showing in Iowa, coupled with
Kemp's tough-talking ads against Bob Dole, could put Kemp in the
campaign say George Bush's poor showing in Iowa, coupled with
Kemp's tough-talking ads against Bob Dole, could put Kemp in the
running for the Republican nomination.
   Before last Monday's Iowa caucuses, Kemp had been on a roll in
New Hampshire, using an effective advertising campaign and the
endorsement of the influential Concord Monitor to help broaden
support.
……
```

Concepts

{N728 refugee, immigrant, migrant},
{N354 friend, colleague, neighbor},
{N118 leader, member, democrat},
{N271 company, industry, business},
{N549 he, I, they},
{N98 clergy, priest, cleric},
{N76 government, authority, administration},
{N561 infringement, encroachment, violation},
{N85 failure, refusal, inability},
{N192 price, rate, amount},
{N289 policy, decision, stance},

Unsupervised Learning from Text


```
<DOC>
<DOCNO> AP880212-0006 </DOCNO>
<FILEID>AP-NR-02-12-88 1644EST</FILEID>
<FIRST>r i AM-CagedHens     02-12 0159</FIRST>
<SECOND>AM-Caged Hens,0162</SECOND>
<HEAD>Court Rules Caging Hens Is Not Cruelty</HEAD>
<DATELINE>STROEMMEN, Norway (AP) </DATELINE>
<TEXT>
   A court ruled Friday that an egg
producer who kept his 2,000 hens in small cages was not guilty of
cruelty to animals, as alleged by animal rights activists.
   ``The verdict is a great relief. It would have been too much to
be found guilty of cruelty to my 2,000 hens,'' Karl Wettre was
quoted as saying by the national NTB news agency after his
acquittal.
   The National Society for the Prevention of Cruelty to Animals
claimed that by keeping hens in small cages, Wettre violated
national legislation to allow animals' natural development and
behavior.
   But the court found that Wettre observed Norwegian regulations
stipulating that a hen should have at least 112 square inches of
cage space in which to live.
   NSPCA chairman Toralf Metveit was quoted as saying: ``I'm
disappointed but not surprised.''
   The society was ordered pay $15,600 in court costs.
</TEXT>
</DOC>
<DOC>
<DOCNO> AP880212-0007 </DOCNO>
<FILEID>AP-NR-02-12-88 1518EST</FILEID>
<FIRST>u p AM-Kemp'sStrategy     02-12 0654</FIRST>
<SECOND>AM-Kemp's Strategy,650</SECOND>
<HEAD>Kemp Strategy To Crack Top Three in N.H.
Primary</HEAD>
<HEAD>With AM-Kemp-Robertson Bjt</HEAD>
<BYLINE>By JONATHAN KELLOGG</BYLINE>
<BYLINE>Associated Press Writer</BYLINE>
<DATELINE>NASHUA, N.H. (AP) </DATELINE>
<TEXT>
   Strategists for Jack Kemp's presidential
campaign say George Bush's poor showing in Iowa, coupled with
Kemp's tough-talking ads against Bob Dole, could put Kemp in the
campaign say George Bush's poor showing in Iowa, coupled with
Kemp's tough-talking ads against Bob Dole, could put Kemp in the
running for the Republican nomination.
   Before last Monday's Iowa caucuses, Kemp had been on a roll in
New Hampshire, using an effective advertising campaign and the
endorsement of the influential Concord Monitor to help broaden
support.
……
```


Unsupervised Learner

Relational Templates

{N728 refugee, immigrant, migrant}, {N271 company, industry, business}, {N549 he, I, they}, …

complained to

{N98 clergy, priest, cleric}, {N76 government, authority, administration}, …

about

{N561 infringement, encroachment, violation}, {N85 failure, refusal, inability}, …

Unsupervised Learning from Text

<DOC>
<DOCNO> AP880212-0006 </DOCNO>
<FILEID>AP-NR-02-12-88 1644EST</FILEID>
<FIRST>r i AM-CagedHens 02-12 0159</FIRST>
<SECOND>AM-Caged Hens,0162</SECOND>
<HEAD>Court Rules Caging Hens Is Not Cruelty</HEAD>
<DATELINE>STROEMMEN, Norway (AP) </DATELINE>
<TEXT>
A court ruled Friday that an egg producer who kept his 2,000 hens in small cages was not guilty of cruelty to animals, as alleged by animal rights activists.
``The verdict is a great relief. It would have been too much to be found guilty of cruelty to my 2,000 hens,'' Karl Wettre was quoted as saying by the national NTB news agency after his acquittal.
The National Society for the Prevention of Cruelty to Animals claimed that by keeping hens in small cages, Wettre violated national legislation to allow animals' natural development and behavior.
But the court found that Wettre observed Norwegian regulations stipulating that a hen should have at least 112 square inches of cage space in which to live.
NSPCA chairman Toralf Metveit was quoted as saying: ``I'm disappointed but not surprised.''
The society was ordered pay $15,600 in court costs.
</TEXT>
</DOC>
<DOC>
<DOCNO> AP880212-0007 </DOCNO>
<FILEID>AP-NR-02-12-88 1518EST</FILEID>
<FIRST>u p AM-Kemp'sStrategy 02-12 0654</FIRST>
<SECOND>AM-Kemp's Strategy,650</SECOND>
<HEAD>Kemp Strategy To Crack Top Three in N.H. Primary</HEAD>
<HEAD>With AM-Kemp-Robertson Bjt</HEAD>
<BYLINE>By JONATHAN KELLOGG</BYLINE>
<BYLINE>Associated Press Writer</BYLINE>
<DATELINE>NASHUA, N.H. (AP) </DATELINE>
<TEXT>
Strategists for Jack Kemp's presidential campaign say George Bush's poor showing in Iowa, coupled with Kemp's tough-talking ads against Bob Dole, could put Kemp in the campaign say George Bush's poor showing in Iowa, coupled with Kemp's tough-talking ads against Bob Dole, could put Kemp in the running for the Republican nomination.
Before last Monday's Iowa caucuses, Kemp had been on a roll in New Hampshire, using an effective advertising campaign and the endorsement of the influential Concord Monitor to help broaden support.
……

Unsupervised Learner

Inference Rules

X complained to Y about Z ≈

X filed a complain about Z with/to Y
X reported Z to Y
a complaint from X about Z
X pleaded with Y
X protested Z
X objected to Z
X decried Z
X is concerned about Z,
….

Outline

- **Distributional Word Similarity**
- Acquisition of Paraphrases
- Clustering By Committee (CBC)
- Relationship to MEANING
- Summary

Distributional Hypothesis

- Words that appear in similar contexts have similar meanings [Harris 69].
- Example: duty vs. responsibility

  -V:from:N absolve 4, back down 1, ban 1, bring 2, Charter 1, come back 2, detach 1, discharge 3, dismiss 1/1, disqualify 1, distance 1, distract 1/2, ease 1, escape 1, excuse 6/1, exempt 3, express 1, flinch 1, free 2/1, get away 1, grow 1, hide 1/1, present 1, reassign 3, release 6/2, relieve 1, remove 17/3, resign 2, retire 10, retreat 1/1, return 11, return home 1, run 1, save 1, separate 1, shield 1, shrink 2, sign off 1, slip away 1, step 1, step down 2, suspect 1, suspend 13, sway 1, take time off 1/1, transfer 1, vary 1
- Demo

Synonyms vs Antonyms (1)

- Example indicators of incompatibility
  - from X to Y
  - either X or Y
- Search results on Alta Vista

adversary NEAR ally	2469	adversary NEAR opponent	2797
“from adversary to ally”	8	“from adversary to opponent”	0
“from ally to adversary”	19	“from opponent to adversary”	0
“either adversary or ally”	1	“either adversary or opponent”	0
“either ally or adversary”	2	“either opponent or adversary”	0

Synonyms vs Antonyms (2)

- Use bilingual dictionaries
  - Obtain potential synonym from other sources unrelated to word distributional.
    - Words with same translation in another language are potentially synonyms.
  - Examples
    - failure → échec, fault → échec
    - path → chemin, thread → chemin
- Intersect them with distributionally similar words

Evaluation

- Data
  - 80 synonyms and 80 antonyms from the Webster's Collegiate Thesaurus that are also top-50 distributionally similar words of each other
- Evaluation task: retrieve synonyms
- Results

Method	Precision	Recall
Pattern-based	86.4	95.0
Bilingual Dictionaries	93.9	39.2

Outline

- Distributional Word Similarity
- **Acquisition of Paraphrases**
- Clustering By Committee (CBC)
- Relationship to MEANING
- Summary

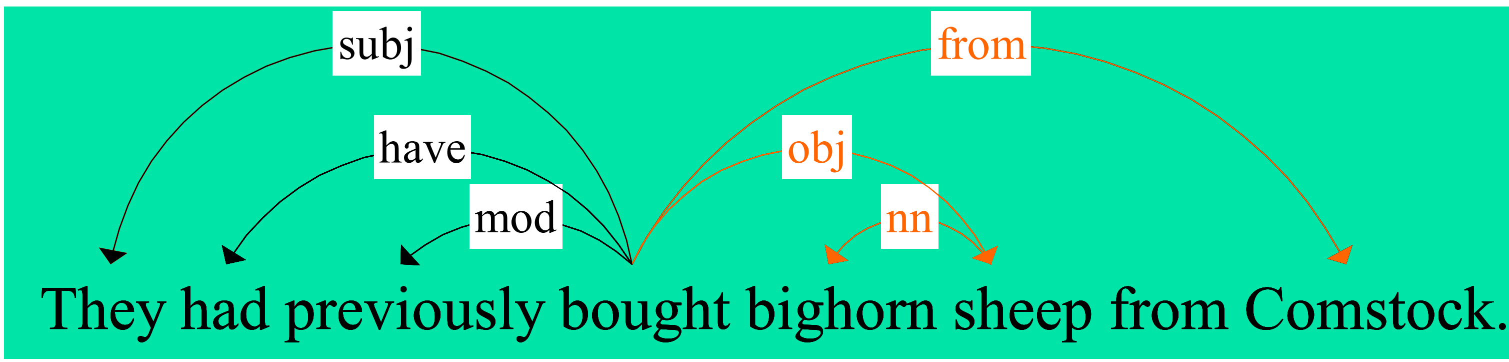
Motivations: Query/Text Mismatch

- Suppose a user asks
  - What does Peugeot manufacture?
- Document may contain:
  - Peugeot is a French car maker;
  - Peugeot builds cars;
  - Peugeot’s production of cars;
  - Peugeot unveils a new compact sedan;
  - Peugeot’s line of minivans;
  - Peugeot’s car factory;
  - ……

Paraphrase: Similar Expressions

- A generalization of similar words.
- Extended Distributional Hypothesis
  - Two expressions are similar if they tend to occur in similar contexts.
- What is an expression?
  - A subtree of a parse tree?
  - A local (one level) tree: X sold Y to Z?
  - A path in a parse tree
    - a binary relationship between two words (nouns).

Paths in Parse Trees



N:from:V<buy>V:obj:N>sheep>N:nn:N

X: Comstock

Y: bighorn

Constraints on Paths

- A path must have at least two links
- A path must begin and end with a noun
- A path must not cross boundaries of finite clauses or adverbial clauses
- All internal links must be frequent
  - OK: N:from:V<buy>V:obj:N>stock>N:nn:N
  - NOT: N:from:V<buy>V:obj:N>sheep>N:nn:N

Similarity between Paths

"X finds a solution to Y"		*"X solves Y"*	
SlotX	***SlotY***	***SlotX***	***SlotY***
commission	**strike**	**committee**	**problem**
committee	**civil war**	**clout**	**crisis**
committee	**crisis**	**government**	**problem**
government	**crisis**	**he**	**mystery**
government	**problem**	**she**	**problem**
he	**problem**	**petition**	**woe**
I	**situation**	**researcher**	**mystery**
legislator	**budget deficit**	**resistance**	**crime**
sheriff	**dispute**	**sheriff**	**murder**

Path similarity is the geometric average of the slot similarities

Experimental Data

- ACQUAINT Data Set (3 GB)
  - Used in TREC Question-Answering Track
  - Contents: AP Newswire, New York Times, Xinghua News (in English)
- Paths extracted:
  - 290M paths (113M unique).
  - 183K paths with frequency counts greater than 50 and total mutual information greater than 300.
- Demo

Limitations

- Synonym vs. Antonym
  - Like other distribution-based learning algorithms, synonyms and antonyms are distributionally indistinguishable.
- Indistinguishable roles
  - When multiple roles of a relations come from the same domain, these roles are indistinguishable.
    - X causes Y

Related Work in Paraphrase Acquisition from Corpus

- From parallel translations of the same novel.
  - Regina Barzilay and Kathleen R. McKeown. (ACL 2001)
- From news stories about the same event.
  - Yusuke Shinyama, Satoshi Sekine, Kiyoshi Sudo and Ralph Grishman. (HLT 2002)
- The documents have to be paraphrases
  - Such data sets are very small.

Outline

- Distributional Word Similarity
- Acquisition of Paraphrases
- **Clustering By Committee (CBC)**
- Relationship to MEANING
- Summary

CBC: A Motivating Example

___ appellate court	campaign in ___
___ capital	governor of ___
___ driver's license	illegal in ___
___ outlaws sth.	primary in ___
___'s sales tax	senator for ___
___'s airport	archbishop of ___
___'s business district	fly to ___
___'s mayor	mayor of ___
___'s subway	outskirts of ___

New York
Washington
California
Texas
Florida
Arizona
Georgia
New Jersey
North Carolina
Iowa
Virginia
Michigan
Massachusetts
New Hampshire
Missouri
Pennsylvania
……

Clustering By Committee (CBC)

- Most clustering algorithms treat each data element as a single point in the feature space.
- Natural language words are often mixture of several points (senses).
- Solution:
  - Define a recruiting committee for each cluster which consists of monosemous words only.

Algorithm

- Phase 1: find top similar words
  - Compute each element's top-*k* most similar elements
- Phase 2: construct committees
  - Find tight clusters among top-k similar words of each given word and use them as candidates for committees.
- Phase 3: create clusters using the committees
  - Similar to K-means

Phase 2: Construct Committees

- Goal: construct committees that
  - form tight clusters (high intra-cluster similarity)
  - dissimilar from other committees (low inter-cluster similarity)
  - cover the whole space
- Method: Find clusters in the top-similar words of every given words

Candidate Committees

```
New York
  __Atlanta 0.18
 |___San Francisco 0.22
 |  |___Chicago 0.23
 |  |  |___Boston 0.26
 |  |___Los Angeles 0.23
 |___New York 0.21
 |___WASHINGTON 0.17
 |___New York City 0.11
Washington
  __San Francisco 0.16
 |___Boston 0.23
 |  |___Chicago 0.26
 |___Los Angeles 0.23
 |___Atlanta 0.22
 |___New York 0.21
 |___Moscow 0.08
    |___Washington 0.18
```

```
California
  __Georgia 0.17
 |___TEXAS 0.13
 |  |___FLORIDA 0.23
 |  |___California 0.21
 |___South Carolina 0.21
Texas
  __Georgia 0.17
 |___ARIZONA 0.14
 |  |___FLORIDA 0.21
 |     |___Texas 0.23
 |___California 0.19
Florida
  __North Carolina 0.14
 |___New Jersey 0.10
 |  |___California 0.14
 |     |___TEXAS 0.21
 |        |___Florida 0.23
 |___Georgia 0.22
```

A Committee and its Features

Committee:

New Delhi
Cairo
Islamabad
Jakarta
Manila
Amman
Seoul

-V:from:N			-N:in:N	
	arrive	9.93	embassy	9.45
	fly	9.76	U.S. Embassy	8.79
	return	7.00	meeting	8.72
	take off	6.95	ambassador	8.54
	travel	6.05	summit	8.45
-V:to:N			-N:gen:N	
	fly	9.67	airport	9.04
	evacuate	7.85	Chinatown	6.78
	send	7.12	district	6.73
	head	6.15	street	6.41
-A:subj:N			-N:mod:A	
	keen	5.50	downtown	8.76
	ready	4.99	capital	7.91
	responsible	3.64	central	7.16

Phase 3: Construct Clusters

- For each word
  - Find its most similar cluster and place the word in the cluster
  - Remove the overlapping features between the word and the cluster
  - Find the next most similar cluster to the residue features
- A word can belong to different clusters
  - Each corresponds to one of its senses.
- Demo

Outline

- Distributional Word Similarity
- Acquisition of Paraphrases
- Clustering By Committee (CBC)
- **Relationship to MEANING**
- Summary

Relationship to MEANING?

- Automatic vs Manual/Semiautomatic Construction of Lexical Knowledge Bases
- Evaluation of Lexical Resources
- Selectional Preference

WordNet is GREAT, but…

- People are very poor at recall
- There are many rare senses
  - almost anything is a person: company, fish, dog, shrimp, ……
- Poor coverage of proper names
  - Nike is a Greek diety

Sample Comparison with WordNet

1	handgun, revolver, shotgun, pistol, rifle, machine gun, sawed-off shotgun, submachine gun, gun, automatic pistol, automatic rifle, firearm, carbine, ammunition, magnum, cartridge, automatic, stopwatch
236	whitefly, pest, aphid, fruit fly, termite, mosquito, cockroach, flea, beetle, killer bee, maggot, predator, mite, houseplant, cricket
471	supervision, discipline, oversight, control, governance, decision making, jurisdiction
706	blend, mix, mixture, combination, juxtaposition, combine, amalgam, sprinkle, synthesis, hybrid, melange
941	employee, client, patient, applicant, tenant, individual, participant, renter, volunteer, recipient, caller, internee, enrollee, giver

Evaluation of Lexical Resources

- Comparison with “Gold Standard”
  - WordNet
  - BBI
  - Roget’s Thesaurus
- Embedded Evaluation: using the resource in an application.
  - Information retrieval
  - Machine translation
  - Language modeling

Color Cluster vs. WordNet

pink, red, turquoise, blue, purple, green, yellow, beige, orange, taupe, color, white, lavender, fuchsia, brown, gray, black, mauve, royal blue, violet, chartreuse, deep red, teal, dark red, aqua, gold, burgundy, lilac, crimson, black and white, garnet, coral, grey, silver, ivory, olive green, cobalt blue, scarlet, tan, amber, cream, rose, indigo, light brown, maroon, uniform, reddish brown, peach, navy blue, plum, nectarine, mulberry, flower, tone, blond, khaki, plaid

Selectional Preference

- Generalization from:
  - drink: beer 151, water 101, alcohol 72, coffee 71, it 62, wine 61, lot 45, milk 28, alcoholic beverage 25, what 24, tea 22, glass 22, more 20, champagne 19, rubbing alcohol 17, bottle 17, ...
- to:
  - drink: {N541 coffee, tea, soft drink} 1289, {N550 whisky, whiskey, cognac} 690, {N592 vinegar, lemon juice, olive oil} 673, {N1358 himself, themselves, myself} 380, {N3 LOT, bit, some} 298, {N792 container, bottle, jar} 203, {N1336 Bud Light, Budweiser, Pepsi} 135, {N949 liqueur, Grand Marnier, brandy} 126, ...

Expectation Maximization

- Generative Model
  - Generate a class for a given context
  - The class generates the word

$$P(c \mid w) = \frac{P(c, w)}{P(w)} = \frac{P(c)P(w \mid c)}{\sum_{c'} P(c')P(w \mid c')}$$

- Problem?
  - The EM model doesn't learn!
  - Solution: learn multiple preferences at the same time.

Summary

- Distinguishing Antonyms from Synonyms
- Paraphrase Acquisition
  - Based on extended distributional hypothesis
  - www.cs.ualberta.ca/~lindek/demos/paraphrase.htm
- Clustering by Committee
  - www.cs.ualberta.ca/~lindek/demos/wordcluster.htm
- Relationship to MEANING
- CYC in a day?

Clustering Similar Paths

N:obj:V<cure>V:subj:N
N:for:N<treatment>N:subj:N
N:obj:V<treat>V:subj:N
N:of:N<variety<N:obj:V<treat>V:subj:N
N:for:N<treatment>N:nn:N
N:for:V<prescribe>V:obj:N
N:obj:V<cure>V:with:N
N:obj:V<diagnose>V:subj:N
N:for:N<therapy>N:nn:N
N:obj:V<treat>V:with:N
N:with:N<patient<N:obj:V<treat>V:subj:N
N:for:V<treat>V:subj:N
N:with:N<people<N:obj:V<help>V:subj:N
N:for:V<prescribe>V:subj:N
N:with:N<people<N:obj:V<treat>V:subj:N
N:obj:V<cure>V:by:N

N:by:N<intervention>N:in:N
N:gen:N<intervention>N:in:N
N:subj:V<intervene>V:in:N
N:nn:N<intervention>N:in:N
N:by:N<interference>N:in:N
N:subj:V<interfere>V:in:N
N:gen:N<interference>N:in:N
N:subj:N<intervention>N:in:N
N:subj:V<meddle>V:in:N
N:subj:V<intervene>V:on:N
N:subj:V<take>V:obj:N>action>N:in:N
N:by:N<intervention>N:nn:N