I256:
Applied Natural Language Processing

Marti Hearst
Nov 13, 2006

Today

Automating Lexicon Construction

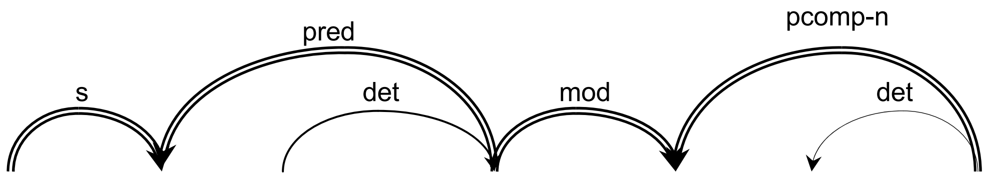
PMI (Turney 2001)

- Pointwise Mutual Information
- Posed as an alternative to LSA
  - score(*choicei*) = log2(p(*problem* & *choicei*) / (p(*problem*)p(*choicei*)))
- With various assumptions, this simplifies to:
  - score(*choicei*) = p(*problem* & *choicei*) / p(*choicei*)
- Conducts experiments with 4 ways to compute this
  - score1(*choicei*) = hits(*problem* AND *choicei*) / hits(*choicei*)

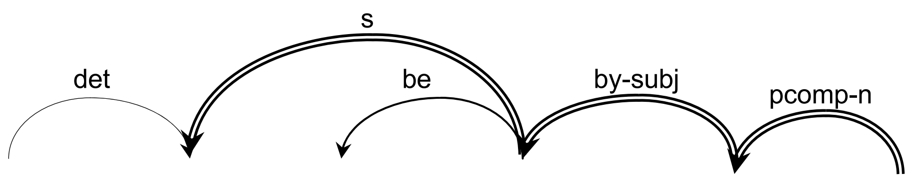
$$\text{score}_4(\mathit{choice}_i) = \frac{\text{hits}((\mathit{problem}\ \text{NEAR}\ \mathit{choice}_i)\ \text{AND}\ \mathit{context}\ \text{AND NOT}\ ((\mathit{problem}\ \text{OR}\ \mathit{choice}_i)\ \text{NEAR "not"}))}{\text{hits}(\mathit{choice}_i\ \text{AND}\ \mathit{context}\ \text{AND NOT}\ (\mathit{choice}_i\ \text{NEAR "not"}))}$$

Dependency Parser (Lin 98)

- Syntactic parser that emphasizes dependancy relationships between lexical items.



- Alice is the author of the book.



- The book is written by Alice

Automating Lexicon Construction

What is a Lexicon?

- A database of the vocabulary of a particular domain (or a language)
- More than a list of words/phrases
- Usually some linguistic information
  - Morphology (manag- e/es/ing/ed → manage)
  - Syntactic patterns (transitivity etc)
- Often some semantic information
  - Is-a hierarchy
  - Synonymy
  - Numbers convert to normal form: Four → 4
  - Date convert to normal form
  - Alternative names convert to explicit form
    - Mr. Carr, Tyler, Presenter → Tyler Carr

Lexica in Text Mining

- Many text mining tasks require named entity recognition.
- Named entity recognition requires a lexicon in most cases.
- Example 1: Question answering
  - Where is Mount Everest?
  - A list of geographic locations increases accuracy
- Example 2: Information extraction
  - Consider scraping book data from amazon.com
  - Template contains field “publisher”
  - A list of publishers increases accuracy
- Manual construction is expensive: 1000s of person hours!
- Sometimes an unstructured inventory is sufficient
- Often you need more structure, e.g., hierarchy

Semantic Relation Detection

- Goal: automatically augment a lexical database
- Many potential relation types:
  - ISA (hypernymy/hyponymy)
  - Part-Of (meronymy)
- Idea: find unambiguous contexts which (nearly) always indicate the relation of interest

Lexico-Syntactic Patterns (Hearst 92)

(S1) `Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.`

(1a) NP_0 *such as* NP_1 $\{, NP_2$ ... , *(and | or)* $NP_i\}$ $\quad i \geq 1$

are such that they imply

(1b) *for all* NP_i, $i \geq 1$, *hyponym*(NP_i, NP_0)

Thus from sentence (S1) we conclude

hyponym("Gelidium", "red algae").

Lexico-Syntactic Patterns (Hearst 92)

(2) *such NP as {NP ,}* {(or | and)} NP*

`...  works by such authors as Herrick, Goldsmith, and Shakespeare.`

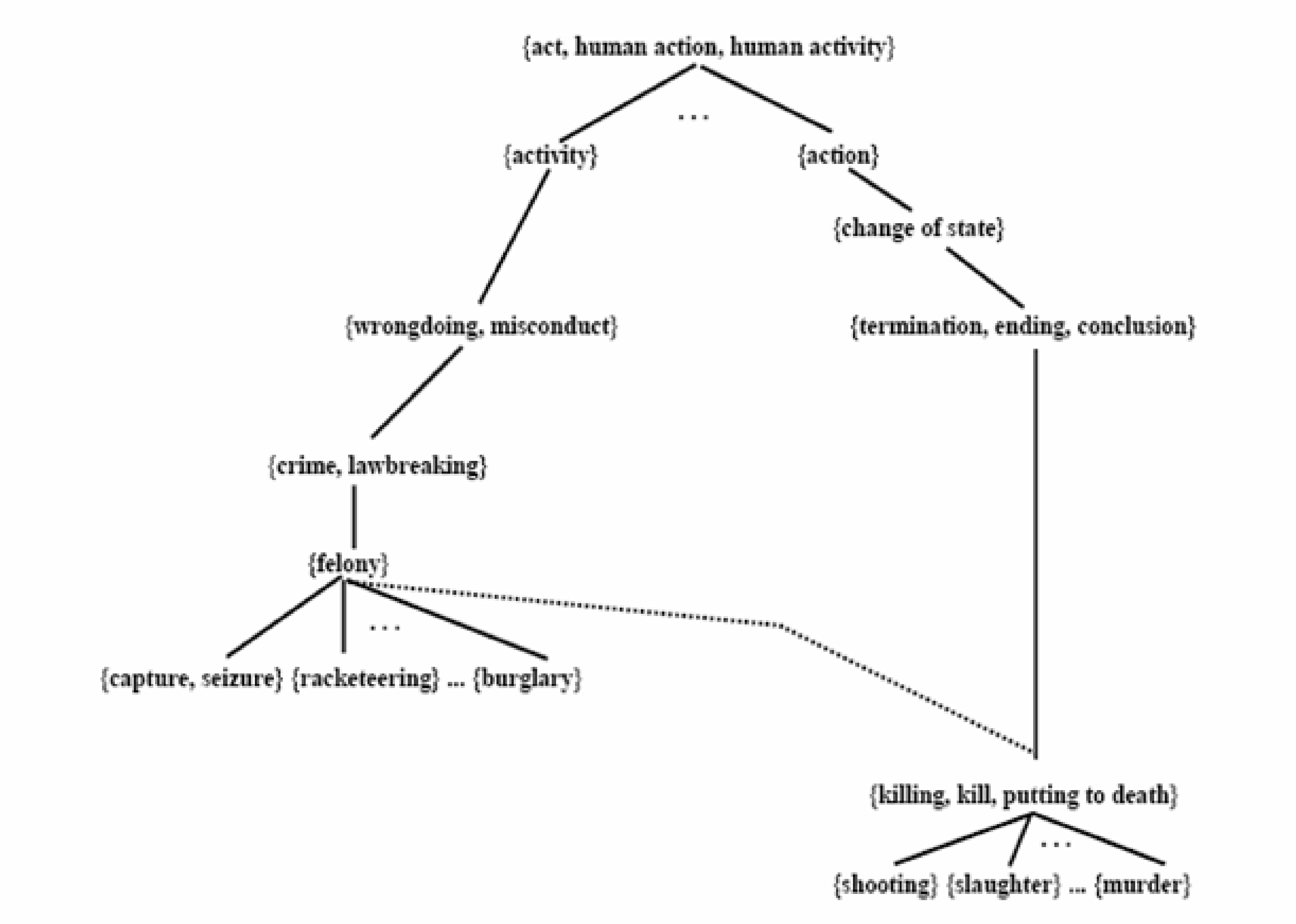
$\Longrightarrow$ *hyponym("author", "Herrick"),*
hyponym("author", "Goldsmith"),
hyponym("author", "Shakespeare")

(3) *NP {, NP}* {,} or other NP*

`Bruises, ..., broken bones or other injuries ...`

$\Longrightarrow$ *hyponym("bruise", "injury"),*
hyponym("broken bone", "injury")

Adding a New Relation



Automating Semantic Relation Detection

- Lexico-syntactic Patterns:
  - Should occur frequently in text
  - Should (nearly) always suggest the relation of interest
  - Should be recognizable with little pre-encoded knowledge.
- These patterns have been used extensively by other researchers.

Lexicon Construction (Riloff 93)

- Attempt 1: Iterative expansion of phrase list
- Start with:
  - Large text corpus
  - List of seed words
- Identify "good" seed word contexts
- Collect close nouns in contexts
- Compute confidence scores for nouns
- Iteratively add high-confidence nouns to seed word list. Go to 2.
- Output: Ranked list of candidates

Lexicon Construction: Example

- Category: weapon
- Seed words: bomb, dynamite, explosives
- Context: <new-phrase> and <seed-phrase>
- Iterate:
  - Context: They use TNT and other explosives.
  - Add word: TNT
- Other words added by algorithm: rockets, bombs, missile, arms, bullets

Lexicon Construction: Attempt 2

- Multilevel bootstrapping (Riloff and Jones 1999)
- Generate two data structures in parallel
  - The lexicon
  - A list of extraction patterns
- Input as before
  - Corpus (not annotated)
  - List of seed words

Multilevel Bootstrapping

- Initial lexicon: seed words
- Level 1: Mutual bootstrapping
  - Extraction patterns are learned from lexicon entries.
  - New lexicon entries are learned from extraction patterns
  - Iterate
- Level 2: Filter lexicon
  - Retain only most reliable lexicon entries
  - Go back to level 1
- 2-level performs better than just level 1.

Scoring of Patterns

- Example
  - Concept: company
  - Pattern: owned by <x>
- Patterns are scored as follows
  - score(pattern) = F/N log(F)
  - F = number of unique lexicon entries produced by the pattern
  - N = total number of unique phrases produced by the pattern
  - Selects for patterns that are
    - Selective (F/N part)
    - Have a high yield (log(F) part)

Scoring of Noun Phrases

- Noun phrases are scored as follows
  - score(NP) = sum_k (1 + 0.01 * score(pattern_k))
  - where we sum over all patterns that fire for NP
  - Main criterion is number of independent patterns that fire for this NP.
  - Give higher score for NPs found by high-confidence patterns.
- Example:
  - New candidate phrase: boeing
  - Occurs in: owned by <x>, sold to <x>, offices of <x>

Shallow Parsing

- Shallow parsing needed
  - For identifying noun phrases and their heads
  - For generating extraction patterns
- For scoring, when are two noun phrases the same?
  - Head phrase matching
  - X matches Y if X is the rightmost substring of Y
  - “New Zealand” matches “Eastern New Zealand”
  - “New Zealand cheese” does not match “New Zealand”

Seed Words

Web Company:	*co. company corp. corporation inc. incorporated limited ltd. plc*
Web Location:	*australia canada china england france germany japan mexico switzerland united_states*
Web Title:	*ceo cfo president vice-president vp*
Terr. Location:	*bolivia city colombia district guatemala honduras neighborhood nicaragua region town*
Terr. Weapon:	*bomb bombs dynamite explosive explosives gun guns rifle rifles tnt*

Mutual Bootstrapping

Generate all candidate extraction patterns from the training corpus using AutoSlog.

Apply the candidate extraction patterns to the training corpus and save the patterns with their extractions to *EPdata*

$SemLex = \{\text{seed\_words}\}$

$Cat\_EPlist = \{\}$

MUTUAL BOOTSTRAPPING LOOP

1. Score all extraction patterns in *EPdata*.
2. *best_EP* = the highest scoring extraction pattern not already in *Cat_EPlist*
3. Add *best_EP* to *Cat_EPlist*
4. Add *best_EP*'s extractions to *SemLex*.
5. Go to step 1

Extraction Patterns

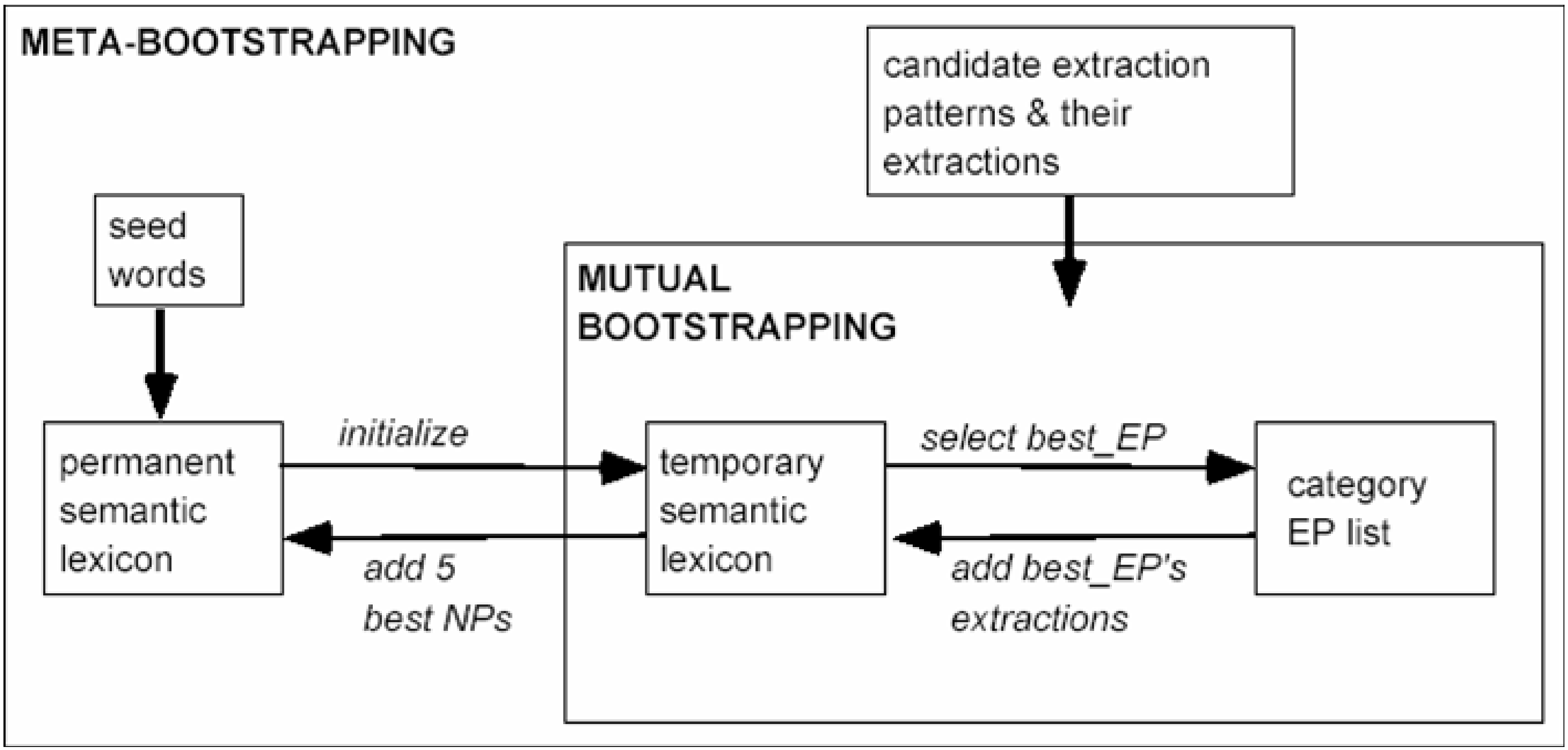
Web Company Patterns
owned by <x>
both as <x>
<x> employed
<x> is distributor
<x> positioning
marks of <x>
motivated <x>
<x> trust company
sold to <x>
devoted to <x>
<x> consolidated stmts.
<x> thrive
message to <x>
<x> is obligations
<x> request information
<x> is foundation
<x> has positions
incorporated as <x>
offices of <x>
<x> required to meet

Level 1: Mutual Bootstrapping

Best pattern	"headquartered in <x>" (F=3,N=4)
Known locations	*nicaragua*
New locations	*san miguel, chapare region, san miguel city*
Best pattern	"gripped <x>" (F=2,N=2)
Known locations	*colombia, guatemala*
New locations	none
Best pattern	"downed in <x>" (F=3,N=6)
Known locations	*nicaragua, san miguel**, *city*
New locations	*area, usulutan region, soyapango*
Best pattern	"to occupy <x>" (F=4,N=6)
Known locations	*nicaragua, town*
New locations	*small country, this northern area, san sebastian neighborhood, private property*
Best pattern	"shot in <x>" (F=5,N=12)
Known locations	*city, soyapango**
New locations	*jauja, central square, head, clash, back, central mountain region, air, villa el_salvador district, northwestern guatemala, left side*

- Drift can occur.
- It only takes one bad apple to spoil the barrel.
- Example: head
- Introduce level 2 bootstrapping to prevent drift.

Level 2: Meta-Bootstrapping



Evaluation

Recall/Precision (%)	*Baseline*	*Lexicon*	*Union*
Web Company	10/32	18/47	18/45
Web Location	11/98	51/77	54/74
Web Title	6/100	46/66	47/62

CoTraining (Collins&Singer 99)

- Similar back and forth between
  - an extraction algorithm and
  - a lexicon
- New: They use word-internal features
  - Is the word all caps? (IBM)
  - Is the word all caps with at least one period? (N.Y.)
  - Non-alphabetic character? (AT&T)
  - The constituent words of the phrase ("Bill" is a feature of the phrase "Bill Clinton")
- Classification formalism: Decision Lists

Collins&Singer: Seed Words

full-string=New_York	→	Location
full-string=California	→	Location
full-string=U.S.	→	Location
contains(Mr.)	→	Person
contains(Incorporated)	→	Organization
full-string=Microsoft	→	Organization
full-string=I.B.M.	→	Organization

Note that categories are more generic than in the case of Riloff/Jones.

Collins&Singer: Algorithm

- Train decision rules on current lexicon (initially: seed words).
  - Result: new set of decision rules.
- Apply decision rules to training set
  - Result: new lexicon
- Repeat

Collins&Singer: Results

Learning Algorithm	Accuracy (Clean)	Accuracy (Noise)
Baseline	45.8%	41.8%
EM	83.1%	75.8%
(Yarowsky 95)	81.3%	74.1%
Yarowsky-cautious	91.2%	83.2%
DL-CoTrain	91.3%	83.3%
CoBoost	91.1%	83.1%

Per-token evaluation?

More Recent Work

- Knowitall system at U Washington
- WebFountain project at IBM

Lexica: Limitations

- Named entity recognition is more than lookup in a list.
- Linguistic variation
  - Manage, manages, managed, managing
- Non-linguistic variation
  - Human gene MYH6 in lexicon, MYH7 in text
- Ambiguity
  - What if a phrase has two different semantic classes?
  - Bioinformatics example: gene/protein metonymy

Discussion

Partial resources often available.

- E.g., you have a gazetteer, you want to extend it to a new geographic area.

- Some manual post-editing necessary for high-quality.
- Semi-automated approaches offer good coverage with much reduced human effort.
- Drift not a problem in practice if there is a human in the loop anyway.
- Approach that can deal with diverse evidence preferable.
- Hand-crafted features (period for "N.Y.") help a lot.