Information Extraction

slides adapted from Jim Martin's
Natural Language Processing class
http://www.cs.colorado.edu/~martin/csci5832/

Motivation for Information Extraction

- When we covered semantic analysis, we focused on
  - The analysis of single sentences
  - A deep approach that could, in principle, be used to extract considerable information from each sentence
  - And a tight coupling with syntactic analysis
- Unfortunately, when released in the wild such approaches have difficulties with
  - Speed... Deep syntactic and semantic analysis of each sentence is too slow for many applications
    - Transaction processing where large amounts of newly encountered text has to be analysed
  - Coverage... Real world texts tend to strain both the syntactic and semantic capabilities of most systems

Information Extraction

- So just as we did with partial parsing and chunking for syntax, we can look for more lightweight techniques that get us most of what we might want in a more robust manner.
  - Figure out the entities (the players, props, instruments, locations, etc. in a text)
  - Figure out how they're related
  - Figure out what they're all up to
  - And do each of those tasks in a loosely-coupled data-driven manner

Targeted Semantic Analysis

- Ordinary newswire text is often used in typical examples.
  - Target ‘who’, ‘what’, ‘when’, ‘where’ of news events
  - And there’s an argument that there are useful applications there
- The real interest/money is in specialized domains
  - Bioinformatics
  - Patent analysis
  - Specific market segments for stock analysis
  - Intelligence analysis
  - Etc.

Overview of Information Extraction

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York

Named Entity Recognition

- Find the named entities and classify them by type.

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Relation Extraction

- Basic task: find all the *classifiable* relations among the named entities in a text (populate a database)…
  - Employs, e.g. { <American, Tim Wagner> }
  - Part-Of, e.g. { <United, UAL>, {American, AMR} >

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York

Event Detection

- Find and classify all the events of interest in a text.
  - Most verbs introduce events/states, but not all (*give* a kiss)
  - Nominalizations often introduce events
    - *Collision, destruction, the running…*

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York

Temporal and Numerical Expressions

- Find all the temporal expressions
  - Normalize them based on some reference point
- Find all the Numerical Expressions
  - Classify by type and Normalize

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York

Template Analysis

- Many news stories have a script-like flavor to them. They have fixed sets of expected events, entities, relations, etc.
- Template, schemas or script processing involves:
  - Recognizing that a story matches a known script
  - Extracting the parts of that script

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York

Information Extraction Typical Tasks

- Named entity recognition and classification
- Coreference analysis
- Temporal and numerical expression analysis
- Event detection and classification
- Relation extraction
- Template analysis

NER (Named Entity Recognition)

- Find and classify all the named entities in a text.
- What is a named entity?
  - A mention of an entity using its name.
    - *Kansas Jayhawks*
  - This is a subset of the possible mentions...
    - *Kansas, Jayhawks, the team, it, they*
- Find means identify the exact span of the mention
- Classify means determine the category of the entity being referred to

Typical NE Types for NewsWire Text

- Some applications add more specific types.

Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes, seas
Geo-Political Entity	GPE	Countries, states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicles	VEH	Planes, trains, and automobiles

Type	Example
People	*Turing* is often considered to be the father of modern computer science.
Organization	The *IPCC* said it is likely that future tropical cyclones will become more intense.
Location	The *Mt. Sanitas* loop hike begins at the base of *Sunshine Canyon*.
Geo-Political Entity	*Palo Alto* is looking at raising the fees for parking in the University Avenue district.
Facility	Drivers were advised to consider either the *Tappan Zee Bridge* or the *Lincoln Tunnel*.
Vehicles	The updated *Mini Cooper* retains its charm and agility.

Ambiguity

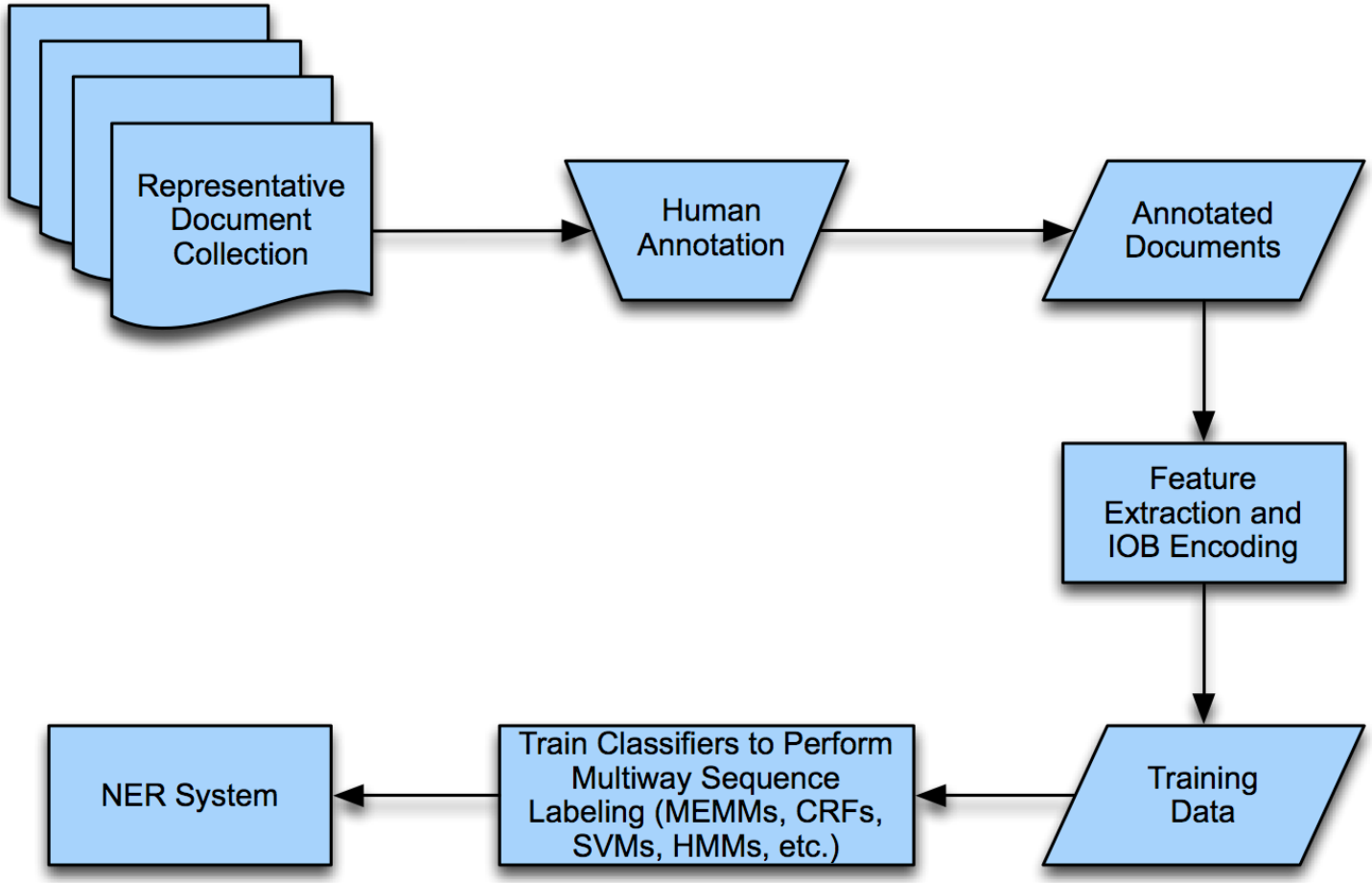
Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Facility
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

[$_{PERS}$ Washington] was born into slavery on the farm of James Burroughs.
[$_{ORG}$ Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [$_{LOC}$ Washington] for what may well be his last state visit.
In June, [$_{GPE}$ Washington] passed a primary seatbelt law.
The [$_{FAC}$ Washington] had proved to be a leaky ship, every passage I made...

NER Approaches

- As with partial parsing and chunking there are two basic approaches (and hybrids)
  - Rule-based (regular expressions)
    - Lists of names
    - Patterns to match things that look like names
    - Patterns to match the environments that classes of names tend to occur in.
  - ML-based approaches
    - Get annotated training data
    - Extract features
    - Train systems to replicate the annotation

ML Approach


Representative Document Collection
Human Annotation
Annotated Documents
Feature Extraction and IOB Encoding
Training Data
Train Classifiers to Perform Multiway Sequence Labeling (MEMMs, CRFs, SVMs, HMMs, etc.)
NER System

Encoding for Sequence Labeling

- Named Entity annotation often uses the IOB encoding:
  - For N classes (i.e. types of entities) we have 2*N+1 tags for tokens
    - B of that class for a token at the beginning of an entity of that class
    - I of that class for a token inside an entity of that class
    - O for a token outside of any class
  - Each token in a text gets a tag.

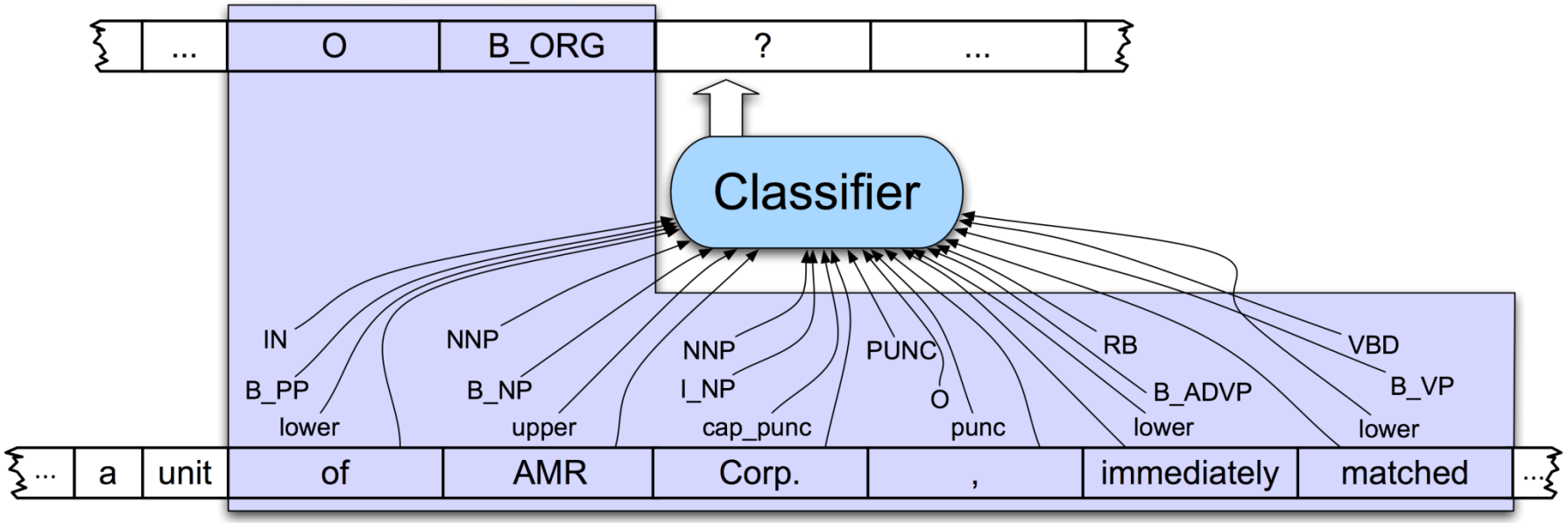
NER Features

- Features may include the word, POS tag, IOB for its phrase type, the shape of the word

Features				Label
American	NNP	B_{NP}	cap	B_{ORG}
Airlines	NNPS	I_{NP}	cap	I_{ORG}
,	PUNC	O	punc	O
a	DT	B_{NP}	lower	O
unit	NN	I_{NP}	lower	O
of	IN	B_{PP}	lower	O
AMR	NNP	B_{NP}	upper	B_{ORG}
Corp.	NNP	I_{NP}	cap_punc	I_{ORG}
,	PUNC	O	punc	O
immediately	RB	B_{ADVP}	lower	O
matched	VBD	B_{VP}	lower	O
the	DT	B_{NP}	lower	O
move	NN	I_{NP}	lower	O
,	PUNC	O	punc	O
spokesman	NN	B_{NP}	lower	O
Tim	NNP	I_{NP}	cap	B_{PER}
Wagner	NNP	I_{NP}	cap	I_{PER}
said	VBD	B_{VP}	lower	O
.	PUNC	O	punc	O

NER as Sequence Labeling

- Classifier may use words and features in from tokens in the sequence before and after the one being classified.



Relations

- Once you have captured the entities in a text you might want to ascertain how they relate to one another, according to the relations of interest
  - Here we're just talking about explicitly stated relations

Relation Extraction

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York

Relation Types

- As with named entities, the list of relations is application specific. For generic news texts...

Relations		Examples	Types
Affiliations			
	Personal	*married to, mother of*	PER → PER
	Organizational	*spokesman for, president of*	PER → ORG
	Artifactual	*owns, invented, produces*	(PER | ORG) → ART
Geospatial			
	Proximity	*near, on outskirts*	LOC → LOC
	Directional	*southeast of*	LOC → LOC
Part-Of			
	Organizational	*a unit of, parent of*	ORG → ORG
	Political	*annexed, acquired*	GPE → GPE

Relations

- By relation we really mean sets of tuples.

Relations	
United is a unit of UAL	$PartOf = \{\langle a,b\rangle, \langle c,d\rangle\}$
American is a unit of AMR	
Tim Wagner works for American Airlines	$OrgAff = \{\langle c,e\rangle\}$
United serves Chicago, Dallas, Denver, and San Francisco	$Serves = \{\langle a,f\rangle, \langle a,g\rangle, \langle a,h\rangle, \langle a,i\rangle\}$

PartOf = {(United, UAL), (American, AMR)}
OrgAff = {(Tim Wagner, American Airlines)}
Serves = {(United, Chicago), (United, Dallas),
(United, Denver), (United, San Francisco)}

Relation Analysis

- As with semantic role labeling we can divide this task into two parts
  - Determining if 2 entities are related
  - And if they are, classifying the relation
- The reason for doing this is two-fold
  - Cutting down on training time for classification by eliminating most pairs
  - Producing separate feature-sets that are appropriate for each relation classification task.

Relation Analysis

- Let's just worry about named entities within the same sentence
  - But, in a system, we will also use entities which are resolved by coreference to pronouns and other referring phrases

```
function FINDRELATIONS(words) returns relations

    relations ← nil
    entities ← FINDENTITIES(words)
    forall entity pairs ⟨e1, e2⟩ in entities do
        if RELATED?(e1, e2)
            relations ← relations+CLASSIFYRELATION(e1, e2)
```

Features

- We can group the features (for both tasks) into three categories
  - Features of the named entities involved
  - Features derived from the words between and around the named entities
  - Features derived from the syntactic environment that governs the two entities

Features

- Features of the entities
  - Their types
    - Concatenation of the types
  - Headwords of the entities
    - *George Washington Bridge*
  - Words in the entities
- Features between and around
  - Particular positions to the left and right of the entities
    - *+/- 1, 2, 3*
    - Bag of words between

Features

- Syntactic environment
  - Constituent path through the tree from one to the other
  - Base syntactic chunk sequence from one to the other
  - Dependency path

Example

- For the following example, we're interested in the possible relation between American Airlines and Tim Wagner.
  - *American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said.*

Entity-based features		
	Entity_1 type	ORG
	Entity_1 head	*airlines*
	Entity_2 type	PERS
	Entity_2 head	*Wagner*
	Concatenated types	ORGPERS
Word-based features		
	Between-entity bag of words	{ *a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman* }
	Word(s) before Entity_1	NONE
	Word(s) after Entity_2	*said*
Syntactic features		
	Constituent path	$NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
	Base syntactic chunk path	$NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
	Typed-dependency path	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

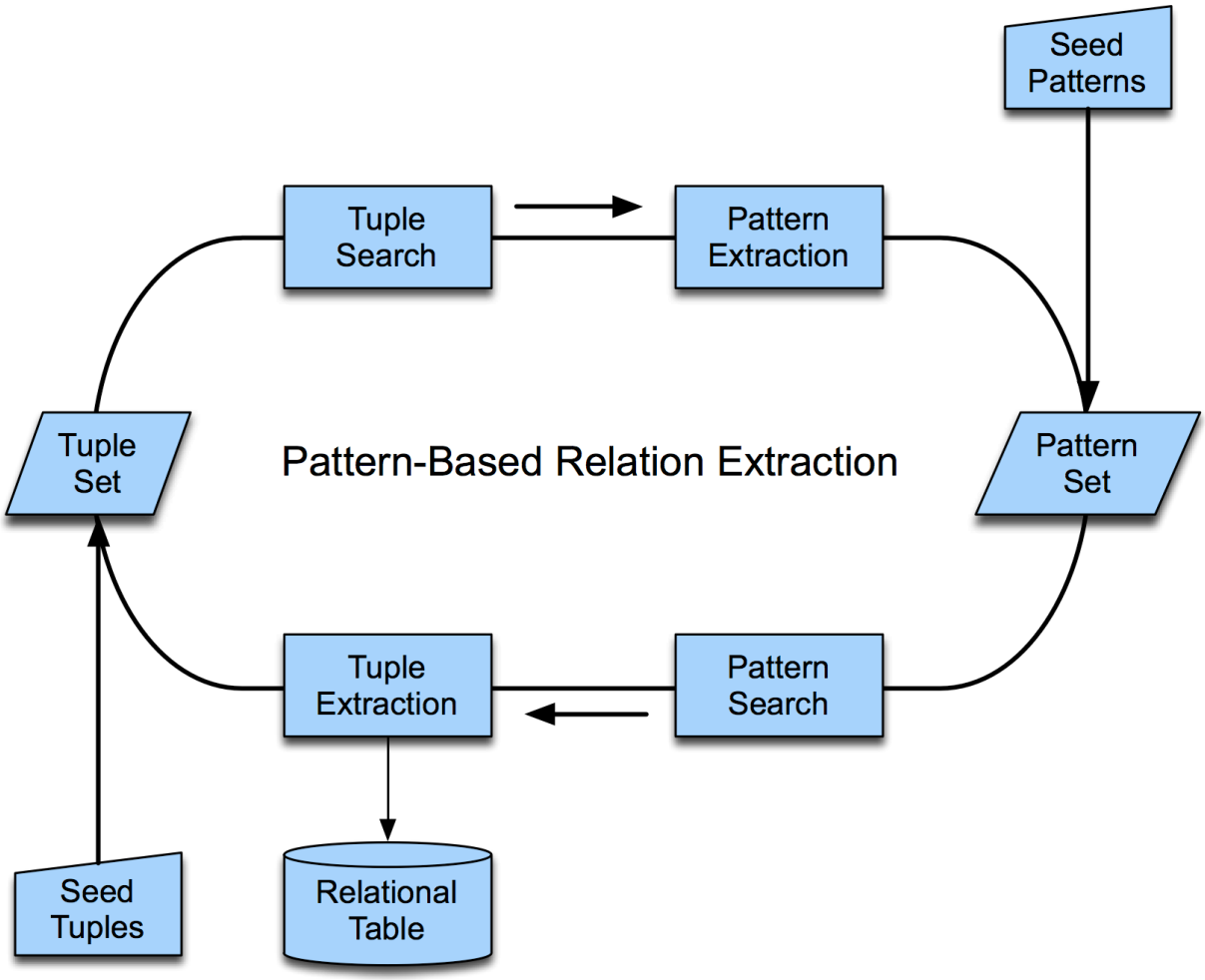
Bootstrapping Approaches

- What if you don't have enough annotated text to train on.
  - But you might have some seed tuples from the annotated text
  - Or you might have some patterns that work pretty well
- Can you use those seeds to do something useful?
  - Co-training and active learning use the seeds to train classifiers to tag more data to train better classifiers...
  - Bootstrapping tries to learn directly (populate a relation) through direct use of the seeds

Bootstrapping Example: Seed Tuple

- <Mark Twain, Elmira> Seed tuple
  - For relation "Location of Burial"
  - Grep (google)
  - "Mark Twain is buried in Elmira, NY."
    - X is buried in Y
  - "The grave of Mark Twain is in Elmira"
    - The grave of X is in Y
  - "Elmira is Mark Twain's final resting place"
    - Y is X's final resting place.
- Use those patterns to google for new tuples that you don't already know

Bootstrapping Relations



Template Filling

- For stories/texts with stereotypical sequences of events, participants, props etc.
- Represent these facts as slots and slot-fillers: templates (frames, scripts, schemas)
  - Evoke the right template
  - Identify the story elements that fill each slot
- Similar approaches as to relation extraction, except that you also have the option of developing patterns or classifiers for more than one slot at once.

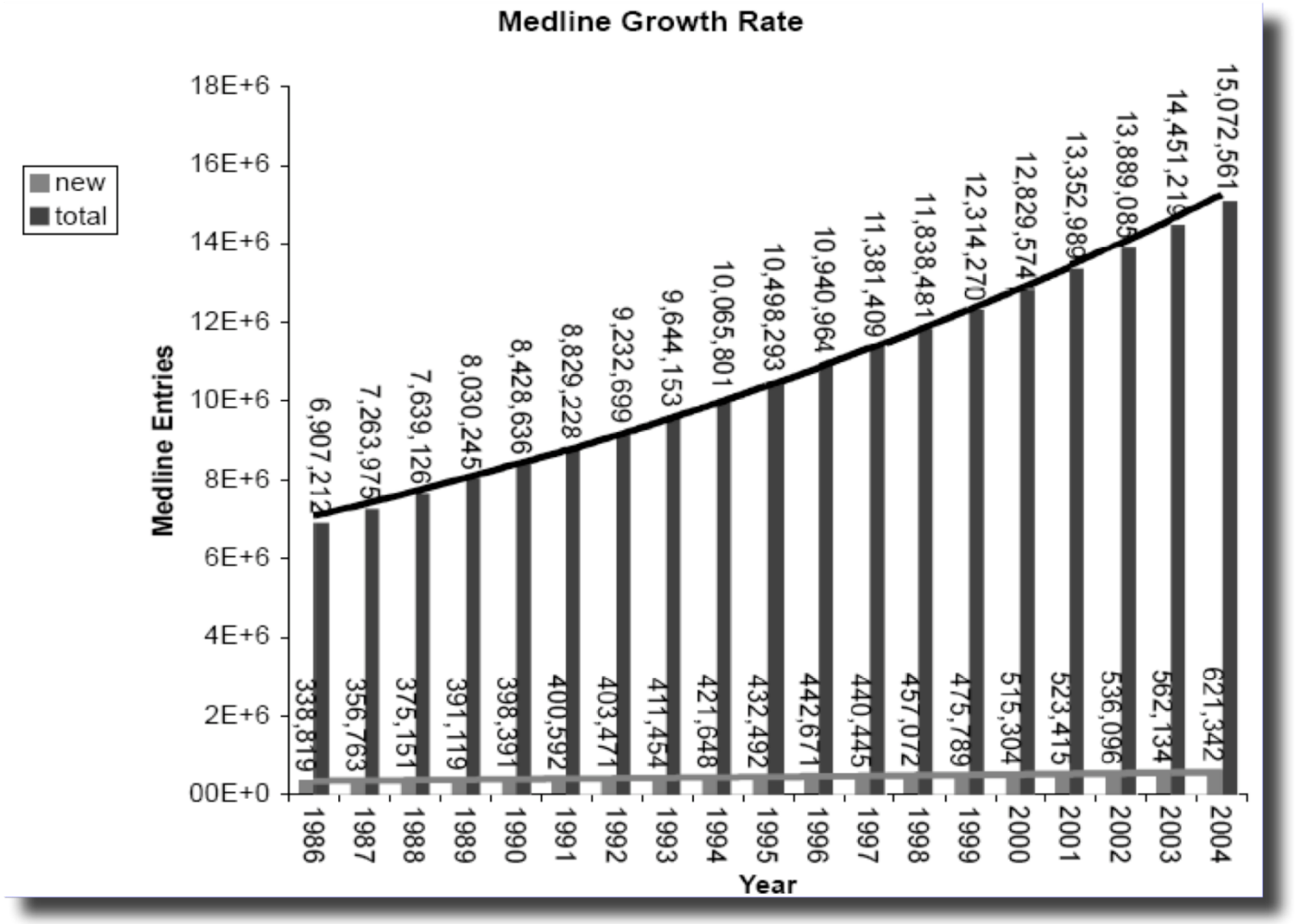
Airline Example

FARE-RAISE ATTEMPT:	LEAD AIRLINE:	UNITED AIRLINES
	AMOUNT:	$6
	EFFECTIVE DATE:	2006-10-26
	FOLLOWER:	AMERICAN AIRLINES

Bioinformatic NLP

- An example domain
  - Very important
  - Practitioners care about the technology
    - They have problems they're trying to solve
  - Lots and lots of text available
  - Lots of interesting problems

Lots of Text


Medline Growth Rate
new
total
Medline Entries
18E+6
16E+6
14E+6
12E+6
10E+6
8E+6
6E+6
4E+6
2E+6
00E+0
6,907,212
7,263,975
7,639,126
8,030,245
8,428,636
8,829,228
9,232,699
9,644,153
10,065,801
10,498,293
10,940,964
11,381,409
11,838,481
12,314,270
12,829,574
13,352,989
13,889,085
14,451,219
15,072,561
338,819
356,763
375,151
391,119
398,391
400,592
403,471
411,454
421,648
432,492
442,671
440,445
457,072
475,789
515,304
523,415
536,096
562,134
621,342
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
Year

Problem Areas

- Mainly variants of NER and relation analysis
  - NER
    - Detecting and classifying named entities
    - And also *normalization*
      - Mapping that named entity to a particular entity in some external database or ontology
  - Relation analysis
    - How various biological entities interact

Bio NER

- Large number of fairly specific types
- Wide (really wide) variation in the naming of entities
  - Gene names
    - *White, insulin, BRCA1, ether a go-go, breast cancer associated 1, etc.*

Semantic class	Examples
Cell lines	*T98G, HeLa cell, Chinese hamster ovary cells, CHO cells*
Cell types	*primary T lymphocytes, natural killer cells, NK cells*
Chemicals	*citric acid, 1,2-diiodopentane, C*
Drugs	*cyclosporin A, CDDP*
Genes/proteins	*white, HSP60, protein kinase C, L23A*
Malignancies	*carcinoma, breast neoplasms*
Medical/clinical concepts	*amyotrophic lateral sclerosis*
Mouse strains	*LAFT, AKR*
Mutations	*C10T, Ala64 → Gly*
Populations	*judo group*

Bio Relations

- Combination of IE and SRL-style relation analysis

(22.27) [*THEME* Full-length cPLA2] was [*TARGET* phosphorylated] stoichiometrically by [*AGENT* p42 mitogen-activated protein (MAP) kinase] in vitro... and the major site of phosphorylation was identified by amino acid sequencing as [*SITE* Ser505]

Bioinformatic IE

- Much work in NLP is concerned with portability and generality
  - How can we get systems trained on one genre/domain to work on a different one
- Biologists don't seem to care much about this...
  - They're happy if you build a specific system to solve their specific problem

Text Analysis Conference (TAC)

- NIST is sponsoring these yearly text analysis tasks (tracks) in the same spirit as TREC for Information Retrieval (IR)
- Knowledge Base Population (KBP)
  - Also tracks on textual entailment and summarization
- Participants must process news articles and prepare an information extraction template formatted as a Wikipedia infobox
  - Must also resolve entities across documents
  - In 2010, must also detect “certainty” of information
  - http://www.nist.gov/tac/