CSCI 5832
Natural Language Processing

Jim Martin

Lecture 22

Today 4/10

- More on IE (Chapter 22)

IE Overview

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

IE Overview

- In case it doesn't become totally obvious...
  - This chapter is just a series of reuses of existing techniques to solve specific problems
    - Partial parsing/chunking
    - Cascades
    - ML sequence labeling
    - Classification/ambiguity resolution

NER

- Find and classify all the named entities in a text.
- What's 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

NE 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

NE Types

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

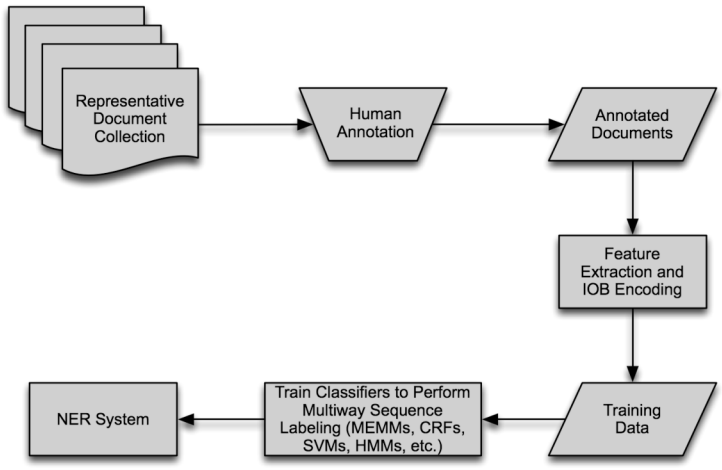
Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Facility
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

[$_{PERS}$ 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



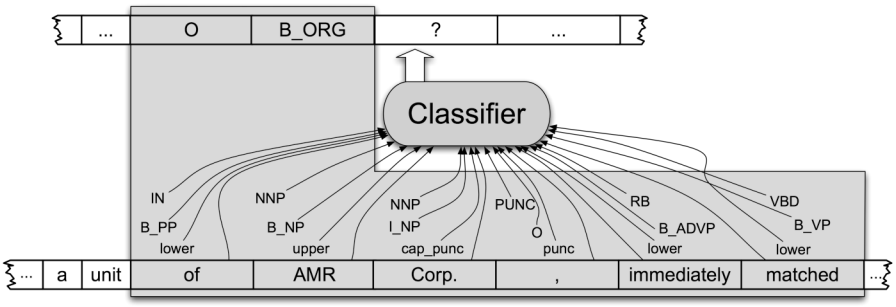
Encoding for Sequence Labeling

- We can use the same IOB encoding here that we used for chunking:
  - For N classes we have 2*N+1 tags
    - An I and B for each class and a O for outside any class.
  - Each token in a text gets a tag.

NER Features

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



NER Evaluation

- As with chunking it is a bad idea to evaluation sequence labelers at the tag level.
  - Most labels are O; so just guessing O gives a learning algorithm a lot of credit.
- So we need to evaluation P/R/F at the entity level.
  - But we may not care equally about all kinds of entities
    - So we might weight them differently in the evaluation routine.

Relations

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

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 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.
  - Think about populating a database.

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	$Serves = \{\langle a,f\rangle, \langle a,g\rangle, \langle a,h\rangle, \langle a,i\rangle\}$

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 task.

Relation Analysis

- Let's just worry about named entities within the same sentence
  - We'll come back to this when we discuss co-reference next week

```
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	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

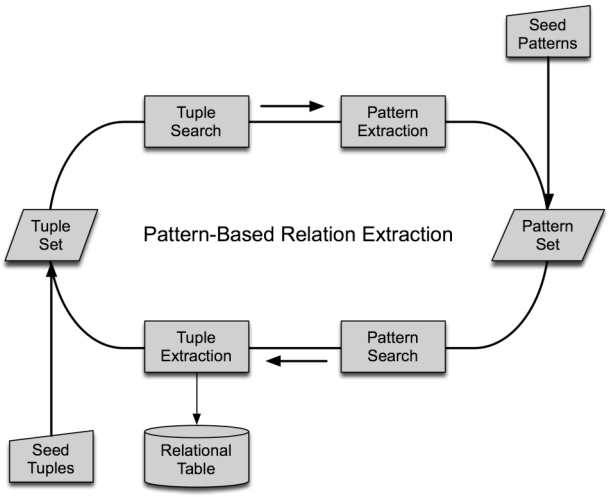
Bootstrapping Approaches

- What if you don't have enough annotated text to train on.
  - But you might have some seed tuples
  - 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
  - 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 grep 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

Airline Example

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

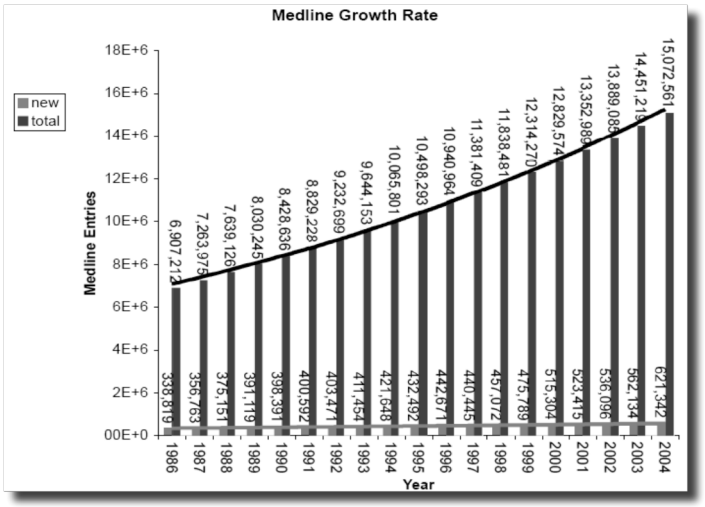
Template-Filling

- Two approaches
  - Cascades of transducers
    - Ala Fastus
  - Supervised ML as Sequence Labeling
    - Two approaches
      - One seq classifier per slot
      - One big sequence classifier

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



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.*

Bio NER Types

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

Next Time

- On (back) to Chapter 21
  - Co-reference
    - Read 21.3 to 21.8
- Quiz is a week from today
  - Covers readings in 17, 18, 20, 21, and 22
    - See schedule page for specific sections
- Final is Monday 5/5 from 1:30 to 4 here in this room