CSCI 5417
Information Retrieval Systems

Jim Martin

Lecture 22
11/10/2011

Today 11/10

- Wrap up Information Extraction
- Start sentiment analysis

IE vs. IR

- Operationally, what IR usually comes down to is the retrieval of documents, not the retrieval of information. It's up to a human to extract the needed information out of the text
- IE is an attempt to automate the extraction of limited kinds of information from free texts
  - These days it's often called *text analytics*
- Sort of sits between NLP and IR

Information Extraction

- So what is it exactly?
  - Figure out the entities (the players, props, instruments, locations, etc. in a text)
  - Figure out how they're related to each other and to other entities
  - Figure out what they're all up to
    - What events they're taking part in
  - And extract information about sentiment and opinion
- And do each of those tasks in a robust, loosely-coupled data-driven manner

IE details

- Going to run through 2 generic applications in more detail
  - NER
  - Relations
- Most other applications are variants on these 2

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

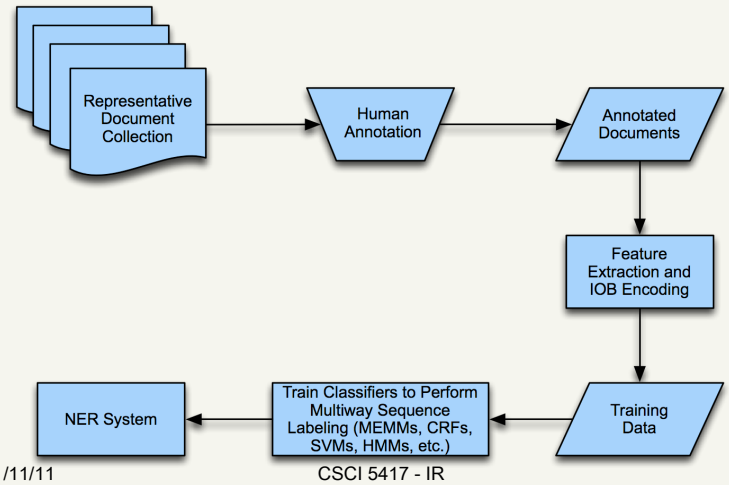
Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Facility
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

[$_{PERS}$ 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 may tasks in IE 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



Data Encoding for Sequence Labeling

- In NER, we're dealing with spans of texts that have been labeled as belonging to some class. So we need to encode
  - The class
  - The start of the span
  - The end of the span
- In a way that is amenable to supervised ML classifiers
  - That is, here's an object represented as a vector of feature/value pairs
  - Here's the class that goes along with that vector

Data Encoding for Sequence Labeling

- The trick with sequences is to come up with an encoding that plays well with the typical classifier
- Popular solution is treat the problem as a word-by-word tagging problem
  - Learn to assign a single tag to each word in a sequence
    - So the tags are the classifier output; the input is some representation of the word in context
  - The tag sequence captures the class, span start, and span finish

IOB Encoding

- A popular way to do this is with IOB encoding. Ignoring classes, every word gets a tag of I (inside), O (outside), or B (begins)

American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said.

B I O O B O O O O O B I O

IOB Encoding

- If we're trying to capture locations, persons, and organizations, we have 3 classes. So we can create, 3 kinds of B and three kinds of I, and leave O as is. That gives is 7 tags.

American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said.

```
B_org   I_org  O O B_org O       O      O  O    O       B_per I_per O
```

In general, for N classes, we wind up with 2*N+1 classes

Training

- So now those tags are the target classifier outputs. We have one object to be classified for each position (token) in the text.
- The features associated with each position are based on
  - Facts based on the word at that position
  - Facts extracted from a window surrounding that position

NE[illegible]es

Word class

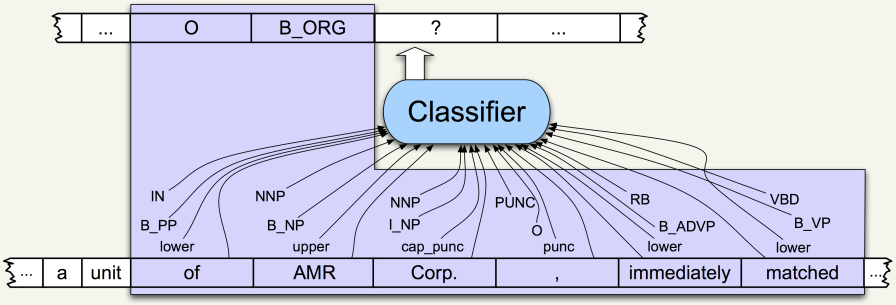
Grammatical chunk

The word itself

Capitalization

Features				Label
American	NNP	B_{NP}	cap	B_{ORG}
Airlines	NNPS	I_{NP}	cap	I_{ORG}
,	PUNC	O	punc	O
[illegible]	DT	B_{NP}	lower	O
[illegible]it	NN	I_{NP}	lower	O
of	IN	B_{PP}	lower	O
AMR	NNP	B_{NP}	upper	$B_{O[illegible]}$
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



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

- 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.

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
  - Information derived from a parse tree for the sentence we're looking at
    - Constituent path through a parse tree from one to the other
    - Base syntactic chunk sequence from one to the other
    - Dependency path

Example

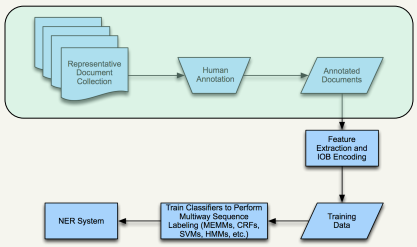
- For the following example, we're interested in the possible relation between American Airlines and Tim Wagner.
  - *American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said.*

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

Relation summary

- Identify entities using NER techniques
- For sentences with multiple entities
  - First detect pairs that are related
  - Then classify the relation
- Both classifiers can trained by extracting features from annotated texts
  - "annotated training data"

Annotation



This is a real bottle-neck to progress.

What about methods to get around this?

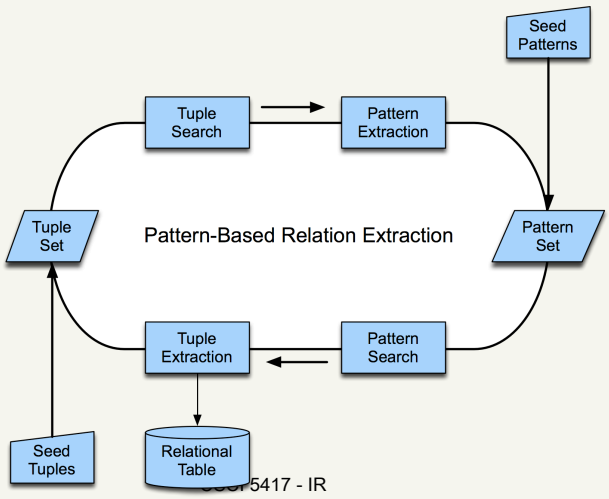
Bootstrapping Approaches

- 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



Problems

- Of course, its not that easy. Unless you want your database populated with non-sense.
- Need to reason about the certainty wrt patterns and tuples.
- If a pattern/tuple is accurate it should be confirmed by evidence from other patterns/tuples
- Constraints on the "types" of the arguments can help filter out false positives

Problems

- So instead of
  - X was buried in Y
  - <person> was buried in <location>

Information Extraction Summary

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

Information Extraction Summary

- All these tasks have two elements in common
  - They involve information explicitly featured in the text
    - Not inferred
  - The information is factual in nature
    - Not subjective
- But of course a lot of what's "in text" is in what isn't said explicitly
  - Text mining
- And a lot of what is said explicitly isn't objective
  - Sentiment analysis

Break

- Quiz is a week from today

Sentiment, Style, Identity, Opinion Classification

- Sentiment Analysis
  - Movie: is a review positive or negative
  - Products (new MacBook Pro)
  - Sentiment over time (is voter anger increasing or decreasing?)
  - Politics (is this editorial left or right?)
  - Prediction (election outcomes, market trends). Will stock go up after this news report?
- Style/Emotion
  - Is this conversation (or blog) friendly, aggressive, polite, flirtatious, deceitful, threatening

Who Cares...

The insurance folks were nice enough to set me up with a rental car until I get my settlement offer. Perfect, since I was planning to rent one to go to Vancouver this weekend anyway, and now its free. They paid for a "standard" size car, which means huge. I asked for something smaller with better fuel economy and ended up with a Kia Rondo in "velvet blue." It is indeed the color of Isabella Rossellini's bathrobe in Blue Velvet.

Every time I drive a rental car I'm a bit appalled. My antique vehicle not only got better gas mileage than most new cars, but it had leg room and head room and ample windows for seeing out. New cars have tiny, low windows with blind spots all over the place. This Kia is ridiculous. It seems to be made for very tall people with very short legs. High ceilings, but the back seat is practically up against the front seat, and the hauling capacity is not better than, say, a Prius.

Example

So what exactly is the point of this compact, yet tall, mid-size SUV? Is it stylish? I can't see any practical reason it is designed this way. It is certainly not an off-road vehicle. I imagine it's front-wheel drive and a bitch to drive in snow. Does simply taking up a lot of space appeal to people? I'm sure it's a fine car, in a general sense, but whatever happened to "smart" design?

Classification

- Coarse-grained classification of sentiment
  - Document-level classification according to some simple (usually binary) scheme
    - Political bias
    - Likes/hates
- Fine-grained classification of sentiment-bearing mentions in a text
  - Positive/negative classification of opinions about entities mentioned in a text
    - Perhaps with intensity

Movies

- A well-studied problem is the problem of classifying movie reviews
  - As either +/-
  - Or on a scale

Movies

- Widely-used corpus available at Cornell
  - http://www.cs.cornell.edu/People/pabo/movie-review-data/

Political Sentiment

- Two examples of classifiers
  - Using words as features
  - And a Naïve Bayes or SVM classifier
  - To make a binary decision
  - About the political stance of a text

Political Sentiment

Bitterlemon.com

- A website designed to "contribute to mutual understanding [between Palestinians and Israelis] through the open exchange of ideas"
- Can we label Israeli & Palestinian perspective:

1. "*The inadvertent killing by Israeli forces of Palestinian civilians – usually in the course of shooting at Palestinian terrorists – is considered no different at the moral and ethical level than the deliberate targeting of Israeli civilians by Palestinian suicide bombers.*"
2. "*In the first weeks of the Intifada, for example, Palestinian public protests and civilian demonstrations were answered brutally by Israel, which killed tens of unarmed protesters.*"

Political Sentiment

- A Naïve Bayes classifier applied to this domain achieved accuracy around 90%

Naïve Bayes: Top 20 words

- Palestinian
  - palestinian, israel, state, politics, peace, international, people, settle, occupation, sharon, right, govern, two, secure, end, conflict, process, side, negotiate
- Israeli
  - israel, palestinian, state, settle, sharon, peace, arafat, arab, politics, two, process, secure, conflict, lead, america, agree, right, gaza, govern

Political Sentiment

- Goal: label a speech as pro or con a bill
- Data: transcripts of all debates in House of Representatives in 2005
  - From GovTrack (http://govtrack.us) website
- Each speech segment (sequence of uninterrupted utterances by speaker)
  - Labeled by the vote ("yea" or "nay") cast
- Labeled by SVM classifier, using all word unigrams as features

Results

- Majority baseline 58.37
- #("support") – #("oppos") 62.67
- SVM classifier 66.05

Choosing a Vocabulary

- Key task: Vocabulary
  - Essentially feature selection
- The previous examples used all words
- Can we do better by focusing on subset of words?
  - How to find words, phrases, patterns that express sentiment or polarity?

Words

- Adjectives
  - positive: honest important mature large patient
    - Ron Paul is the only honest man in Washington.
    - Kitchell's writing is unbelievably mature and is only likely to get better.
    - To humour me my patient father agrees yet again to my choice of film

Words

- Adjectives
  - negative: harmful hypocritical inefficient insecure
    - It was a macabre and hypocritical circus.
    - Why are they being so inefficient ?

Other parts of speech

- Verbs
  - positive**: praise, love**
  - negative**: blame, criticize**
- Nouns
  - positive**: pleasure, enjoyment**
  - negative**: pain, criticism**

Phrases

- Phrases containing adjectives and adverbs
  - positive: high intelligence, low cost
  - negative: little variation, many troubles

Discussion

- What makes classification hard?
  - Sentiment can be subtle:
    - Perfume review in "Perfumes: the Guide":
      - "If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut."
    - "She runs the gamut of emotions from A to B"
      (Dorothy Parker on Katherine Hepburn)
- Order effects
  - This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.