CSCI 5417
Information Retrieval Systems

Jim Martin

Lecture 21
11/8/2011

Today

- Finish learning to rank
- Information extraction

Machine Learning for ad hoc IR

- We've looked at methods for ranking documents in IR using factors like
  - Cosine similarity, inverse document frequency, pivoted document length normalization, Pagerank, etc.
- We've looked at methods for classifying documents using supervised machine learning classifiers
  - Naïve Bayes, kNN, SVMs
- Surely we can also use such machine learning to rank the documents displayed in search results?

Why is There a Need for ML?

- Traditional ranking functions in IR used a very small number of features
  - Term frequency
  - Inverse document frequency
  - Document length
- It was easy to tune weighting coefficients by hand
  - And people did
    - But you saw how "easy" it was on HW1

Why is There a Need for ML

- Modern systems – especially on the Web – use a large number of features:
  - Log frequency of query word in anchor text
  - Query term proximity
  - Query word in color on page?
  - # of images on page
  - # of (out) links on page
  - PageRank of page?
  - URL length?
  - URL contains "~"?
  - Page edit recency?
  - Page length?
- The *New York Times* (2008-06-03) quoted Amit Singhal as saying Google was using over 200 such features.

Using ML for ad hoc IR (Approach 1)

- Well classification seems like a good place to start
  - Take an object and put it in a class
    - With some confidence
  - What do we have to work with in terms of training data?
    - Documents
    - Queries
    - Relevance judgements

Training data

example	docID	query	cosine score	ω	judgment
Φ_1	37	linux operating system	0.032	3	*relevant*
Φ_2	37	penguin logo	0.02	4	*nonrelevant*
Φ_3	238	operating system	0.043	2	*relevant*
Φ_4	238	runtime environment	0.004	2	*nonrelevant*
Φ_5	1741	kernel layer	0.022	3	*relevant*
Φ_6	2094	device driver	0.03	2	*relevant*
Φ_7	3191	device driver	0.027	5	*nonrelevant*

Using classification for ad hoc IR

- A linear scoring function on these two features is then

  $Score(d, q) = Score(\alpha, \omega) = a\alpha + b\omega + c$

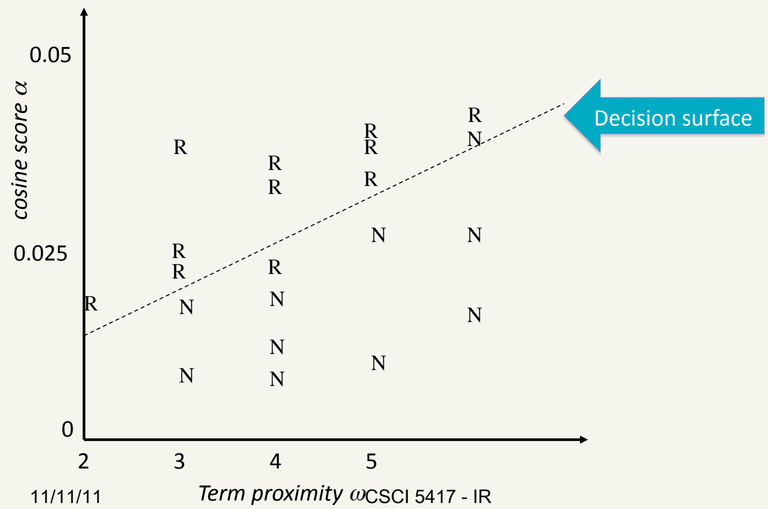
- And the linear classifier is

  Decide relevant if $Score(d, q) > \theta$

- … just like when we were doing text classification

Sec. 15.4.1

Using classification for ad hoc IR



More Complex Cases

- We can generalize this to classifier functions over more features
- We can use any method we have for learning the linear classifier weights

Problem

- The ranking in this approach is based on the classifier's confidence in its judgment
- It's not clear that that should directly determine a ranking between two documents
  - That is, it gives a ranking of confidence not a ranking of relevance
    - Maybe they correlate, maybe not

Learning to Rank

- Maybe classification isn't the right way to think about approaching ad hoc IR via ML
- Background ML
  - Classification problems
    - Map to a discrete unordered set of classes
  - Regression problems
    - Map to a real value
  - Ordinal regression problems
    - Map to an *ordered* set of classes

Learning to Rank

- Assume documents can be totally ordered by relevance given a query
  - These are totally ordered: $d_1 < d_2 < \ldots < d_J$
  - This is the ordinal regression setup
- Assume training data is available consisting of document-query pairs represented as feature vectors ψ_i and a relevance ranking between them

- Such an ordering can be cast as a set of ***pair-wise*** **judgements**, where the input is a pair of results for a single query, and the class is the relevance ordering relationship between them

11/11/11 CSCI 5417 - IR 13

Learning to Rank

- But assuming a total ordering across all docs is a lot to expect
  - Think of all the training data
- So instead assume a smaller number of categories ***C*** of relevance exist
  - These are totally ordered: $c_1 < c_2 < \ldots < c_J$
    - Definitely rel, relevant, partially, not relevant, really really not relevant... Etc.
    - Indifferent to differences within a category
- Assume training data is available consisting of document-query pairs represented as feature vectors ψ_i and relevance ranking based on the categories ***C***

11/11/11 CSCI 5417 - IR 14

The Ranking SVM

[Herbrich et al. 1999, 2000; Joachims et al. 2002]

- Aim is to classify instance pairs as correctly ranked or incorrectly ranked
  - This turns an ordinal regression problem back into a binary classification problem
- We want a ranking function f such that

$$c_i > c_k \text{ iff } f(\psi_i) > f(\psi_k)$$

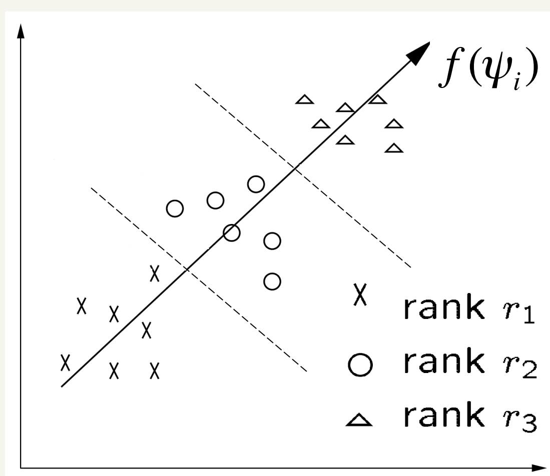
- Suppose that f is a linear function

$$f(\psi_i) = \mathbf{w} \bullet \psi_i$$

The Ranking SVM

[Herbrich et al. 1999, 2000; Joachims et al. 2002]

- Ranking Model: $f(\psi_i)$



The Ranking SVM

[Herbrich et al. 1999, 2000; Joachims et al. 2002]

- Then

$$c_i > c_k \text{ iff } \mathbf{w}\bullet(\psi_i - \psi_k) > 0$$

- So let's directly create a new instance space from such pairs:

$$\Phi_u = \Phi(d_i, d_j, q) = \psi_i - \psi_k$$

$$z_u = +1, 0, -1 \text{ as } c_i >,=,< c_k$$

- From training data $S = \{\Phi_u\}$, we train an SVM

Limitations of Machine Learning

- Everything that we have looked at (and most work in this area) produces *linear* models of features by weighting different base features
- This contrasts with most of the clever ideas of traditional IR, which are *nonlinear* scalings and combinations of basic measurements
  - log term frequency, idf, pivoted length normalization
- At present, ML is good at weighting features, but not at coming up with nonlinear scalings
  - Designing the basic features that give good signals for ranking remains the domain of human creativity

Break

- Quiz 2 Readings (IIR)
  - 13: Skip 13.3, 13.4, 13.5
  - 14: Skip 14.4, 14.5, 14.6
  - 15: Skip 15.2
  - 16: Skip 16.4.1, 16.5
  - 17: Skip 17.5, 17.8
  - 19: All
  - 20: All
  - 21: All

Break

- Quiz 2 Readings (additional)
  - Topic model paper
  - Information extraction chapter
  - Sentiment book readings
    - TBA

IE vs. IR

- Remember from the first class...

Information Retrieval

Information retrieval is the science of searching for information in documents, searching for documents themselves, searching for metadata which describe documents, or searching within databases, whether relational stand-alone databases or hypertextually-networked databases such as the World Wide Web.

Wikipedia

Finding material of an unstructured nature that satisfies an information need from within large collections.

Manning et al 2008

The study of methods and structures used to represent and access information.

Witten et al

The IR definition can be found in this book.

Salton

IR deals with the representation, storage, organization of, and access to information items.

Salton

Information retrieval is the term conventionally, though somewhat inaccurately, applied to the type of activity discussed in this volume.

van Rijsbergen

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

Why

- If you can transform unstructured information found in texts to structured database-like information... You can
  - Improve retrieval of relevant information
  - Enable data-intensive analytics
    - Data-mining, business intelligence, predictive tools, etc.
  - Direct question answering

Web

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

Information Extraction

- Ordinary newswire text is often used in typical examples.
  - And there's an argument that there are useful applications in that domain
- But the real interest/money is in specialized domains
  - Bioinformatics
  - Patent analysis
  - Specific market segments for stock analysis
  - SEC filings
  - Intelligence analysis
  - Health
    - Electronic medical records

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

Information Extraction: Entities

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.

Named Entity Recognition

- Find the named entities and classify them by type.
- Typical approach
  - Acquire training data
  - Train a system with supervised ML
  - Augment with pre- and post-processing using available list resources (census data, gazeteers, etc.)

Information Extraction: Relations

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 (i.e., populate a database)…
  - Employs
    - { <American, Tim Wagner> }
  - Part-Of
    - { <United, UAL>, {American, AMR} >

Relation Extraction

- Typical approach:
  For all pairs of entities in a text
  - Extract features from the text span that just covers both of the entities
    - Use a binary classifier to decide if there is likely to be a relation
    - If yes: then apply each of the known classifiers to the pair to decide which one it is
- Use supervised ML to train the required classifiers from an annotated corpus

Information Extraction: Events

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

Event Detection

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

Information Extraction: Times, numbers, measures, etc.

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

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

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

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

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

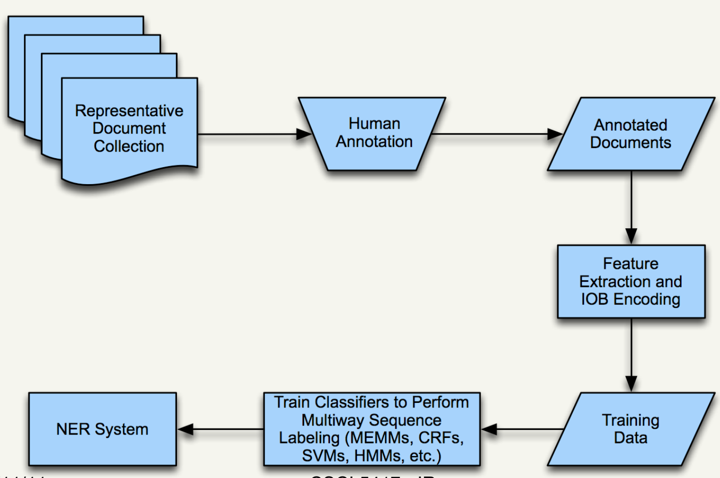
Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Facility
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

[$_{PERS}$ 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

NER Features

Word class

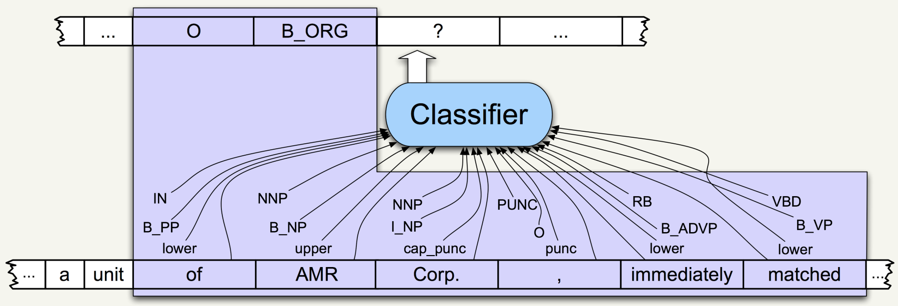
Grammatical chunk

The word itself

Capitalization

Features				Label
American	NNP	B_{NP}	cap	B_{ORG}
Airlines	NNPS	I_{NP}	cap	I_{ORG}
,	PUNC	O	punc	O
	DT	B_{NP}	lower	O
it	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



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

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

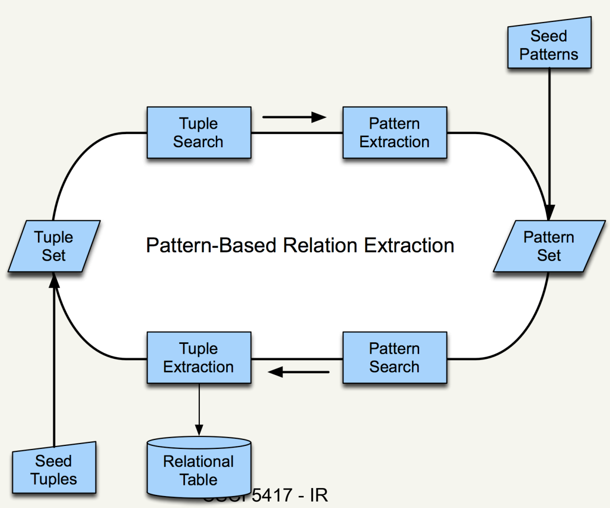
Bootstrapping Approaches

- 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



Information Extraction Summary

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

Social Media and IR

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?