Natural Language Processing

Lecture 20 – 11/3/2015
Jim Martin

Today

- HW 2
- Information extraction
- Briefly review sequence labeling and POS tagging
  - HMMs & MEMMs
- More information extraction

Assignment 2

- Apply naïve Bayes to a sentiment task
  - Hotel reviews
- Due next Thursday (11/12)
- Postponing the quiz

Assignment 2

- I mistakenly thought that since my multiple stays at other Marriott locations were excellent and clean that this location would be as well. Man was I wrong. This place is terrible. The room was very dirty and there were dead bugs everywhere. The room smelled horrible. I couldn't get the room windows open enough to help remove the smell. It took a while to find this place with all the construction in the area and the awkward road design in the area. I will never stay here again!
- I took a trip to New York for a couple of days with my two daughters and we decided to splurge and stay at the Doubletree in Times Square. We booked a small suite on a mid level floor and were very excited by our room. The decor was exciting, bright colored and contemporary. Everything was comfortable and the bathroom was somewhat luxorious. We tried the exercize room too. It was quite nice and we were the only people in it. The reception area was stylish and appealing and the hotel staff were helpful and freindly. No one was snobby even though we were paying over $400 per night. We all had a great time at the Doubletree and felt welcome and comfortable.

Assignment 2

- NB boils down to training a unigram language model for the two classes.
- Classifying a document is just computing P(C|D) for each of the two classes and taking the argmax
  - Product of (x | c) ; where x is a word feature
    - Sum of the logprobs
- For this assignment, assume the classes are equally likely (ignore the class prior)

Assignment 2

- 636 the
- 491 and
- 357 a
- 284 was
- 264 to
- 246 The
- 230 I
- 210 in
- 188 of
- 165 is

- 655 the
- 353 and
- 324 to
- 304 was
- 276 I
- 257 a
- 187 The
- 186 in
- 166 of
- 158 room

Information Extraction

- Ordinary newswire text is often used in typical examples.
  - And there are lots of useful applications out there
- But the real interest/money is in specialized domains
  - Bioinformatics
  - Electronic medical records
  - Stock market analysis
  - Intelligence analysis
  - Social media

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

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

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 A

and New Yor

Organizations	People	Places
United Airlines	Tim Wagner	Chicago
American Airlines		Dallas
AMR		Atlanta
UAL		Denver
		San Francisco
		Los Angeles
		New York

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

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

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

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

Statistical Sequence Labeling

- We can treat NER as a per word tagging task
- Recall with POS tagging we trained systems to tag words using annotated training data
- Training data
  - Hand tag a bunch of data with POS tags
- Training
  - HMMs
  - Logistic regression

HMM Tagging

- Same as we did with POS tagging
  - Argmax P(T|W) = P(W|T)P(T)
  - The tags are the hidden states
- Works ok, but has one significant shortcoming
  - The typical kinds of things that we might think would be useful in this task aren't easily squeezed into the HMM model
- We'd like to be able to make arbitrary features available for the statistical inference being made.
- For that we'll turn to classifiers created using classical machine learning techniques

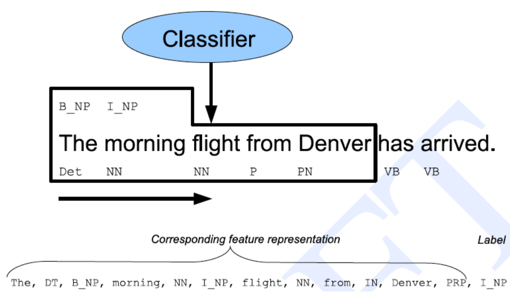
From Classification to Sequence Processing

- Applying classifiers to tagging…
  - The object to be tagged is a word in the sequence
  - The features are
    - features of the word,
    - features of its immediate neighbors,
    - and features derived from the entire context
  - Sequential tagging means sweeping a classifier across the input assigning tags to words as you proceed.

Typical Features

- Typical setup involves
  - A small sliding window around the object being tagged
  - Features extracted from the window
    - Current word token
    - Previous/next N word tokens
    - Current word POS
    - Previous/next POS
    - Capitalization information
    - ...

Statistical Sequence Labeling



Problem

- We're making a long series of local judgments. Without attending to the overall goodness of the final sequence of tags.
- Just hoping that local conditions will yield global goodness.
- HMMs don't have this problem since the language model worried about the overall goodness of the tag sequence.
  - But we don't want to use HMMs since we can't easily squeeze arbitrary features into the learning framework

Answer

- Graft a language model onto the sequential classification scheme.
  - Instead of having the classifier emit one label as an answer for each object, get it to emit a distribution over the labels for each word
  - Train a language model for the kinds of sequences we're trying to produce.
  - Run Viterbi over the label distributions for the sequence to get the best overall sequence

11/3/15 Speech and Language Processing - Jurafsky and Martin 22

MEMMs

- Maximum Entropy Markov Models are one way to do this.
  - Although people do the same thing in an ad hoc way with other classifiers
- MEMMs combine two techniques
  - Logistic regression-based classifiers for the individual labeling
  - Markov models for the sequence model.

11/3/15 Speech and Language Processing - Jurafsky and Martin 23

MaxEnt

$$p(c|x) = \frac{1}{Z}\exp\sum_i w_i f_i$$

11/3/15 Speech and Language Processing - Jurafsky and Martin 24

MaxEnt

$$p(c|x) = \frac{\exp\left(\sum_{i=0}^{N} w_{ci} f_i\right)}{\sum_{c' \in C} \exp\left(\sum_{i=0}^{N} w_{c'i} f_i\right)}$$

Features

$$f_3(c,x) = \begin{cases} 1 & \text{if suffix}(word_i) = \text{"ing"} \ \& \ c = \text{VBG} \\ 0 & \text{otherwise} \end{cases}$$

$$f_4(c,x) = \begin{cases} 1 & \text{if is\_lower\_case}(word_i) \ \& \ c = \text{VB} \\ 0 & \text{otherwise} \end{cases}$$

Features

$$f_3(c,x) = \begin{cases} 1 & \text{if suffix}(word_i) = \text{"ing"} \ \& \ c = \text{VBG} \\ 0 & \text{otherwise} \end{cases}$$

$$f_4(c,x) = \begin{cases} 1 & \text{if is\_lower\_case}(word_i) \ \& \ c = \text{VB} \\ 0 & \text{otherwise} \end{cases}$$

- Key point. You can't squeeze features like these into an HMM.

Back to Sequences

HMMs

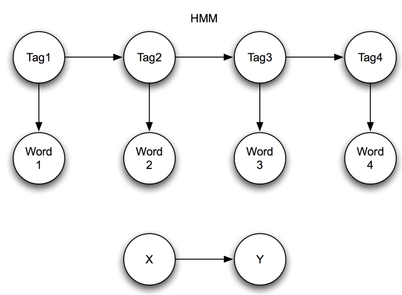
$$\hat{T} = \underset{T}{\operatorname{argmax}} P(T|W)$$
$$= \underset{T}{\operatorname{argmax}} P(W|T)P(T)$$
$$= \underset{T}{\operatorname{argmax}} \prod_i P(word_i|tag_i) \prod_i P(tag_i|tag_{i-1})$$

- MEMMs

$$\hat{T} = \underset{T}{\operatorname{argmax}} P(T|W)$$
$$= \underset{T}{\operatorname{argmax}} \prod_i P(tag_i|word_i, tag_{i-1})$$

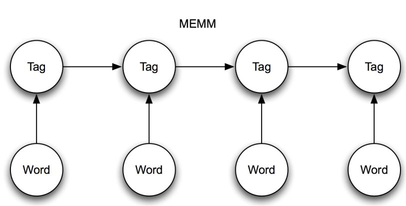
And whatever other features you choose to use!

HMMs vs. MEMMs



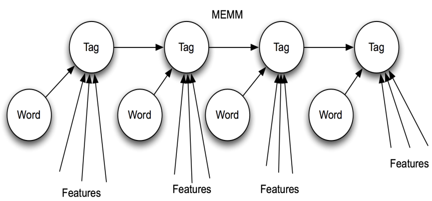
means P(Y|X)

HMMs vs. MEMMs



$$P(T|W) = \prod P(t_i|t_{i-1}, w_i)$$

HMMs vs. MEMMs



$$P(T|W) = \prod P(t_i|t_{i-1}, w_i, f_i)$$

Spans to Tags

- So how do we use word by word tagging to solve the problem of search for spans of text?
- We'll use what's generically called per word IOB encoding
  - I -> Inside ; this word is inside a span
  - O -> Outside ; outside a span of interest
  - B -> Begin ; begins a span

IOB Encoding (Syntax)

The morning flight from Denver has arrived.
B_NP I_NP I_NP O B_NP O O

- This example shows the encoding if we were just looking noun phrases.

The morning flight from Denver has arrived
B_NP I_NP I_NP B_PP B_NP B_VP I_VP

- This example shows full coverage. In this scheme there are 2*N+1 tags. Where N is the number of constituents in your set.

IOB Encoding (NER)

- For each kind of entity, we'll have a specific I and and B tag
  - B_loc, B_person, B_protein, B_org…
- And one general O tag
- Giving is 2*N + 1 kinds of tags

- Tags are the labels that a supervised learner has to learn to emit on a per word basis

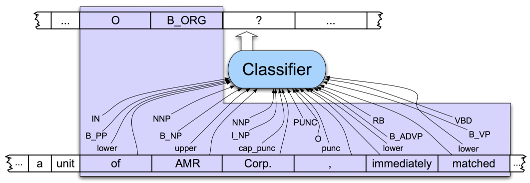
NER Features

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 Features

- The usefulness of different features varies by domain and by language
- But features should be superficial and easily extracted from the text to be analyzed
  - Can't solve a problem by using a feature that's harder to extract than the actual problem!
- The "shape" feature turns out to be amazingly useful in many domains

NER as Sequence Labeling



NER Evaluation

- 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 evaluate precision, recall and F at the entity level.
  - But we may not care equally about all kinds of entities
    - So we might weight them differently in our evaluation.

NER and Entities

- Traditionally, NER only refers to entities that are referred to with an explicit mention of a name.
  - “Jane Smith” vs. “she”
  - “Twitter” vs. “it” or “they”
  - “Tesla Model S” vs. “the car”
- General entity reference and tracking is a bigger problem.

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

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	Arg 1	Arg 2
PartOf	United	UAL
PartOf	American	AMR
EmployedBy	Tim Wagner	American

Relation Analysis

- We can divide relation analysis into two parts
  - Determining if 2 entities are related
  - And if they are, classifying the relation
- There are 2 reasons to do this
  - 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

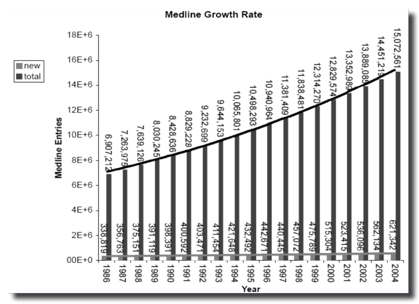
- For the following example, we're interested in the possible relation between American Airlines and Tim Wagner.
  - *American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said.*

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

Case Study: Bioinformatics

- An example domain
  - Very important: basic science, clinical practice, insurance billing, etc.
  - Practitioners care about the technology
    - They have problems they're trying to solve
  - Lots and lots of text available
  - Lots of interesting problems

Lots and 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 quite insane) 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*

Summary

- Information extraction makes use of loosely coupled systems to extract shallow semantic elements from texts
- Must exploit domain dependent features to get state of the art performance
- Current research focused on less objective characteristics of text
  - sentiment, opinion, deception, bias, motivation, predatory intent, etc.