NLP for the Web

Lecture 5
Web Genre Identification and Sequence Tagging

Prof. Dr. Chris Biemann
Seid Muhie Yimam, MA

Readings for This Lecture

For this week:

Mandatory: http://nlp.stanford.edu/IR-book/

- Chapter 13.1-13.2 (pages 253-262): Text Classification and Naïve Bayes
- Chapter 13.4 (pages 253-262): Feature Selection
- Chapter 13.5 (pages 279-284): Text Classification Evaluation

Optional:

Dillon, A. and Gushrowski, B. (2000) Genres and the Web - is the home page the first digital genre? Journal of the American Society for Information Science, 51(2), 202-205

Today's lecture

- Web Genre
  - basics about genres and web genres
  - characteristics of web genres

- Machine Learning:
  Genre Identification as Text Categorization
- Machine Learning:
  Sequence Tagging

Document Types and Genres

- One of the important characteristics of any **document type** is the role it plays in supporting a discourse community

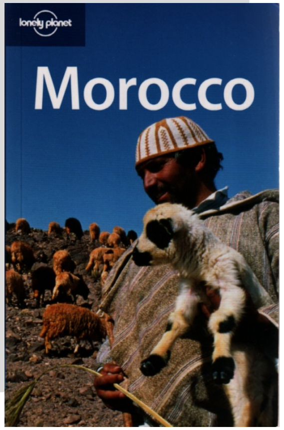
- **Document types evolve** over a long time (decades or even centuries) of use to constitute highly conventional forms that are recognized as being of a **specific type or genre**

- **A genre is:**
  - from French "kind" or "sort", from Latin: *genus* (stem *gener-*)
  - a loose set of criteria for a category of composition
  - often used to categorize literature and speech

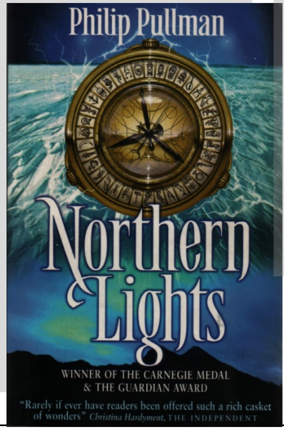
There is no standard inventory of genres in the **web context**.

Genres: Dimensions and Examples

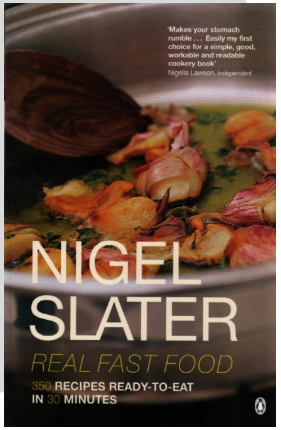
- Genre examples: detective stories, scientific articles, newspapers, catalogs …. , shopping list, flyer, project proposal, etc.
- The characteristic properties of a discourse genre:
  - **Content** topic
  - **Form** layout, design, text structure etc.
  - **Function** communicative purpose etc.



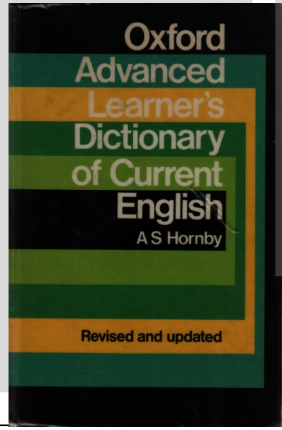
Guidebook



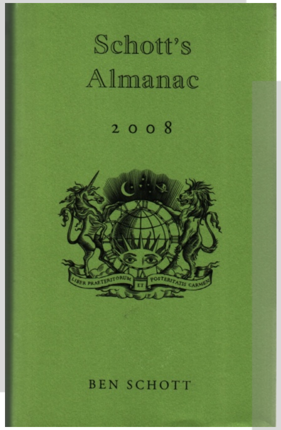
Novel



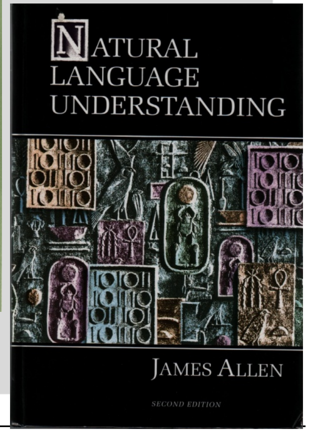
Cookbook



Dictionary



Almanac



Textbook

Motivation for Genres

- Genres serve a **specific community**: Reiffel (1999) describes how the highly stylized form of mathematics writing serves the community of scholars in this domain
- Genre conventions enhance **memorability of discourse** (van Dijk and Kintsch, 1983)
- Genre conventions lead to greater **user satisfaction** (Bazerman, 1988)
- The **genre form** is highly tied to the **behavior of its users** (Reiffel, 1999)
- User orientation and navigation in the information space is **contingent** on the **user's perception of genre** conventions
- A **lack** of genre conventions in the digital world is a potentially significant **source of user difficulty**

Bazerman, C. (1988) Shaping Written Knowledge. The Genre and Activity of the Experimental Article in Science. Madison WI: University of Wisconsin Press.

van Dijk, T.A. and Kintsch, W. (1983) Strategies of Discourse Comprehension. London: Academic Press

Reiffel, E. (1999) The genre of mathematics writing and it's implications for digital documents. In Proc. of the 32nd Annual Hawaii International Conference on System Sciences. Los Alamitos, CA: IEEE Computer Society (published on CD-ROM)

Web Genres

- **Adoption** of many **existing** paper-based conventions: familiarity of form leverages user comprehension
- Web versions of such paper formats as newspapers and magazines frequently **adhere closely** to this type
  - to guide interaction
  - to locate sections of interest
  - to browse the available information

- IT technologies enable new affordances. Therefore, utilize the power of the new medium to provide **innovative information structures**
- Merely inheriting genre conventions from the paper world may be of **disadvantage** to support adequate design of new information types

Definition: Web Genre

Definition Web Genre:

Information spaces that do not have paper equivalents on which they may be modeled yet which manifest genre properties of conventional form, features and organization (A. Dillon and B. Gushrowski, 2000)

- First unique web genre: personal homepage
- The web is still relatively new, so it is NOT quite clear how to apply traditional notions of genre to web pages

Dillon, A. and Vaughan, M. (1997) "It's the journey and the destination": shape and the emergent property of genre in evaluating digital documents. New Review of Hypermedia & Multimedia, 3, 91-106.

Dillon, A. and Gushrowski, B. (2000) Genres and the Web - is the home page the first digital genre? Journal of the American Society for Information Science, 51(2), 202-205

Genres and Corpus-Based Research

- Many corpora used for NLP research, but:
  - very few large corpora **indicate genres**
  - when they do, the typology of genres varies widely

- For example:
  - the Brown corpus uses 15 textual categories added after the corpus construction, from press reportage (a text genre) to religion or skills and hobbies (domains)
  - the British National Corpus (BNC) uses 70 classes, such as academic or non-academic scientific texts or biography, also written vs. spoken

- The genre attribute included in a few collections used in Information Retrieval (TREC HARD 2003 & 2004)

Applications for Genres in NLP

- **Information Retrieval**
  - Filter out irrelevant documents returned by keywords
  - Keywords mostly express the topic of a document (e.g. politics, sports, football, finance, etc.)
  - Genre expresses the type of the text (e.g. newspaper article, technical report, PhD thesis, weather report, etc.)
    - **textbook** web pages that contain **“Dijkstra algorithm”**
    - **PhD thesis** web pages that contain **“web genres”**

- **Information Extraction**
  - Identify & extract useful relevant content from web pages
  - Use genre and layout characteristics

Applications for Genres in NLP and vice versa

- **Information Science**
  - Automatic extraction of metadata for better management and use of digital documents
  - filter in field-based search masks (e.g. library book search)

- **Accuracy of NLP tools** can be increased if targeted for genres
  - Certain types of entities (newspaper vs. biomedical papers)
  - Words and word meanings (tagging, WSD)
  - Certain constructions occur only in certain types of text (parsing)

- **Accuracy of Genre Identification** can be increased with NLP-based features

Web Genre Category Sets in the Literature

(Meyer zu Eissen and Stein, 2004)	Help; Article; Discussion; Shop; Portrayal (non-private); Portrayal (private); Link Collection; Download
(Lim et al., 2005)	Personal homepages; Public homepages; Commercial homepages; Bulletin collections; Link collections; Image collections; Simple tables/lists; Input pages; Journalistic materials; Research reports; Official materials; Informative materials; FAQs; Discussions; Product specifications; Others
(Stubbe et al., 2007a)	Journalism (Commentary; Review; Portrait; Marginal Note; Interview; News; Feature Story; Reportage); Literature (Poem; Prose; Drama); Information (Science Report; Explanation; Recipe; FAQ; Lexicon; Word List; Bilingual Dictionary; Presentation; Statistics; Code); Documentation (Law; Official Report; Protocol); Directory (Person; Catalog; Resources; Timeline); Communcation (Mail/Talk; Forum; Blog; Form); Nothing
(Vidulin et al., 2007)	Pornographic; Blog; Childrens'; Commercial/Promotional; Community; Content Delivery; Entertainment; Error Message; FAQ; Gateway; Index; Informative; Journalistic; Official; Personal; Poetry; Prose Fiction; Scientific; Shopping; User Input
(Braslavski, 2007)	Official, academic, journalistic, literary, and everyday communication style

Georg Rehm, Marina Santini, Alexander Mehler, Pavel Braslavski, Rüdiger Gleim, Andrea Stubbe, Svetlana Symonenko, Mirko Tavosanis, Vedrana Vidulin: Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems. Proceedings of LREC 2008, Marrakech, Morocco

Characteristics of Web Genres

- **Higher complexity** in comparison with traditional genres:
  - Hypertext links
  - Interactive features
  - Multimedia
  - Web 2.0 – Elements

- **Examples** of web genres:
  - Personal homepage
  - FAQ
  - Blog
  - Search engine
  - Encyclopedia
  - Web shop

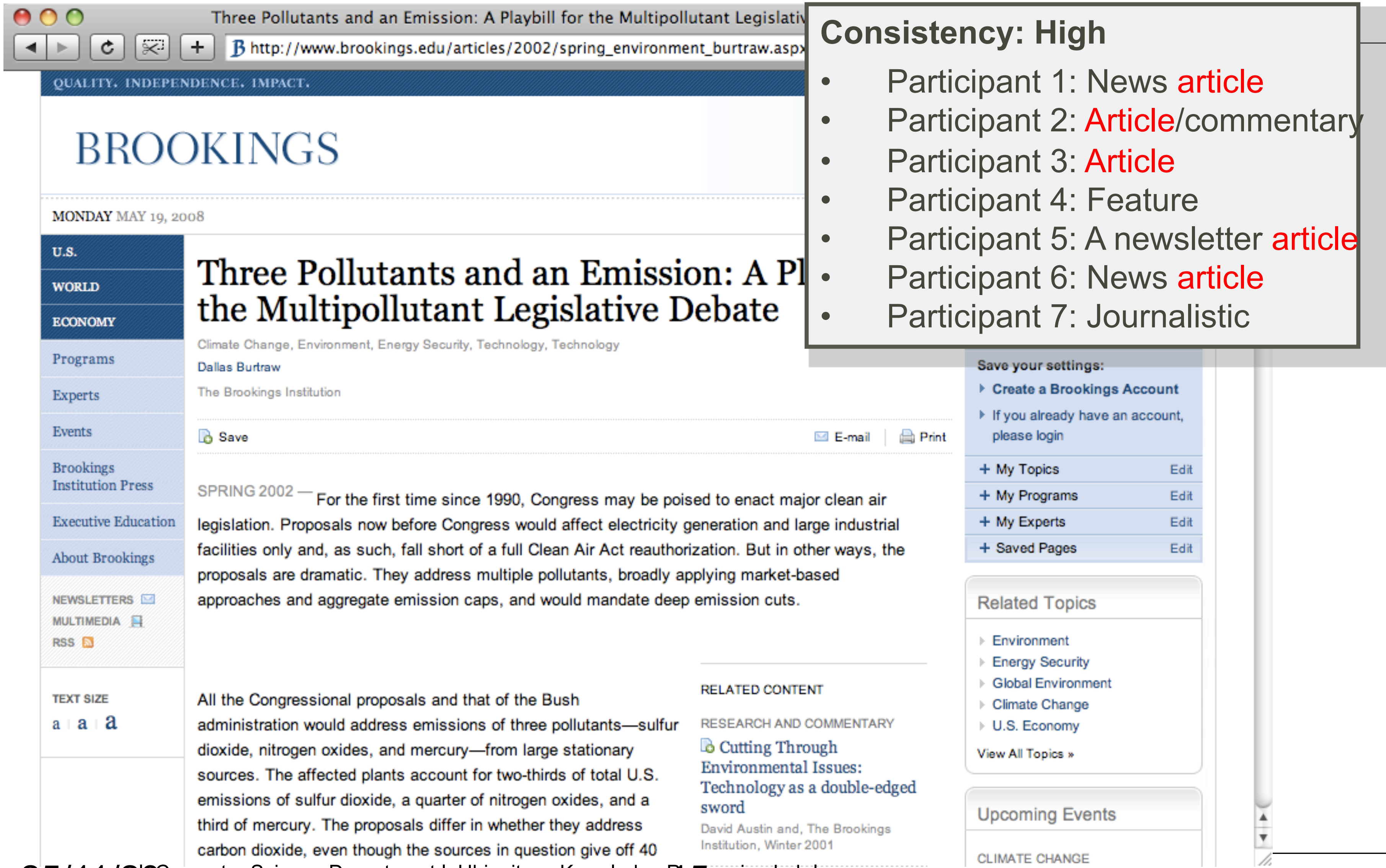
Corpus-Based Genre Study (Rehm et al. 2008)

- **Task:** assigning genre labels to web documents

- **Experimental setup:**
  - 50 randomly selected web pages
  - 7 annotators noted their genre labels in a spreadsheet
  - NO guidelines to assign genre labels
  - Multi-labeling allowed

- **Finding:**
- A high number of disparate labels, from genres and super-genres to descriptions, functional or purpose-oriented properties of documents, even topical categories

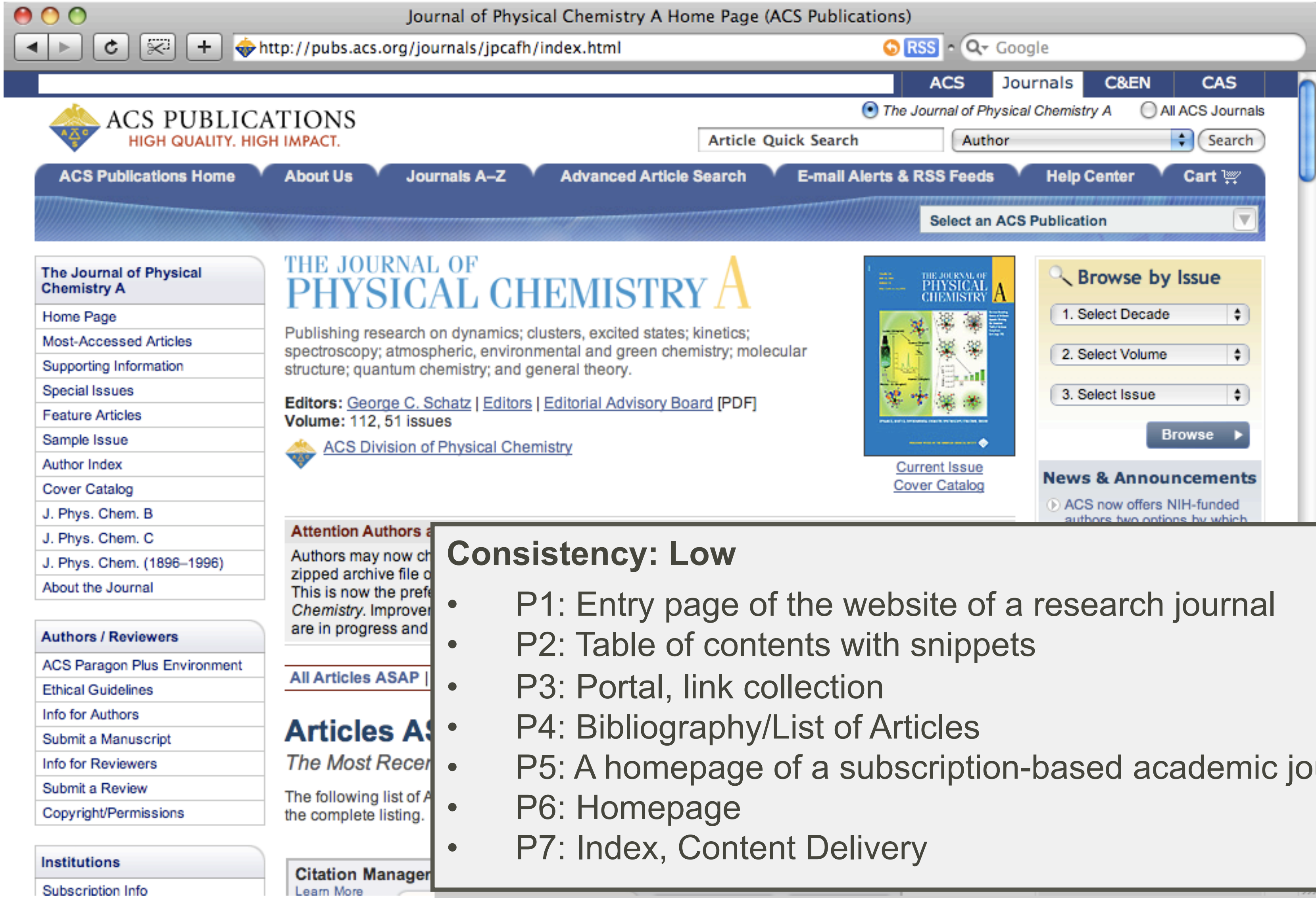
Assigning Genre Labels to Web Pages



Consistency: High

- Participant 1: News article
- Participant 2: Article/commentary
- Participant 3: Article
- Participant 4: Feature
- Participant 5: A newsletter article
- Participant 6: News article
- Participant 7: Journalistic

Assigning Genre Labels to Web Pages



Consistency: Low

- P1: Entry page of the website of a research journal
- P2: Table of contents with snippets
- P3: Portal, link collection
- P4: Bibliography/List of Articles
- P5: A homepage of a subscription-based academic journal
- P6: Homepage
- P7: Index, Content Delivery

Summary of Explorative Annotation Study

Consistency	No. of annotators who used same genre label	No. of documents	
High	5 to 7	6	12%
Medium	3 or 4	26	52%
Low	1 or 2	18	36%

- **Problem:** different level of abstraction or generalization, e.g. „article“, „review“, „reportage“, „a new product“, or „journalistic“ for one web page

- Still, a certain level of agreement exists for **familiar genres** with very large discourse communities, e.g. *blog*, *academic article*, *newspaper article*
- **Conclusions**:
  - assigning genre labels to web documents is hard
  - clear annotation guidelines to specify "ground-truth" are crucial

Automatic Web Genre Identification

- Frequent **classification methods**: Naïve Bayes (NB) & Support Vector Machines (SVM)
- **Classification features** employed:
  - function words (or simply the most frequent words)
  - punctuation
  - POS trigrams
  - trigrams of function words only (all other words are represented by their POS codes or as NON-FUNCTION)
  - more complex syntactic features (e.g. the number of *that* clauses)
  - links to other pages with similar properties
  - page structure and *html* (e.g. length, headlines, lists, avg. line length)
  - specific wordlists (e.g. names, cities, keywords for programming languages)
  - specific POS (e.g. positive ADJ, female pronouns)
  - non-textual items: dates, ordinal numbers, numbers, images, emoticons
  - language models

Supervised Machine Learning and NLP Features

Needed:

- A working definition of “web genre”
- An experimental document collection
- A set of web genre labels
- A corpus annotated with these labels

Machine Learning Setup:

- Learn a classifier function from feature representations to labels
- Train the classifier on a training set
- Test the classifier on a test set

Feature Sources:

- by external knowledge
- by its appearance: surface features
- by a preprocessing step
- by surrounding items
- by smaller items it is composed of
- by combining several features

Value ranges:

- binary
- numeric
- nominal

Example: knowledge-based features

	list of cities	list of countries	combine	database
WORD	**Is-CITY?**	**Is-COUNTRY?**	**Is-CITYor COUNTRY?**	**Inhabitants**
Darmstadt	TRUE	FALSE	TRUE	142K
is	FALSE	FALSE	FALSE	-1
located	FALSE	FALSE	FALSE	-1
in	FALSE	FALSE	FALSE	-1
GERMANY	FALSE	TRUE	TRUE	80M
.	FALSE	FALSE	FALSE	-1

Example: surface features

feature function

IDENTITY FUNCTION

String cap (String word) {.. }

boolean[] cap (String word) {.. }

word.substr(length(word)-2,2)

log(length(word))

WORD	WORD	CAP	CAP-IC	CAP-LC	CAP-AC	CAP-NL	last-2	loglength
Darmstadt	Darmstadt	IC	TRUE	FALSE	FALSE	FALSE	dt	2.197
is	is	LC	FALSE	TRUE	FALSE	FALSE	is	0.6931
located	located	LC	FALSE	TRUE	FALSE	FALSE	ed	1.9459
in	in	LC	FALSE	TRUE	FALSE	FALSE	in	0.6931
GERMANY	GERMANY	AC	FALSE	FALSE	TRUE	FALSE	NY	1.9459
.	.	NL	FALSE	FALSE	FALSE	TRUE	.	0

Example: Features by preprocessing steps

Tree Tagger		NER	Dep. Parser	POS-of head

WORD	POS	lemma	BIO-NER	HEAD POS
Darmstadt	NNP	Darmstadt	B-LOC	VBN
is	VBZ	are	O	VBN
located	VBN	locate	O	root
in	IN	in	O	NNP
GERMANY	NNP	GERMANY	B-LOC	VBN
.	$.	.	O	root

Example: by surrounding items

W: t-1

P: t-1 P: t-2

W: t-1 W: t-2 list of patterns

WORD	POS	WORD-1	POS-1,-2	is-located-pattern
Darmstadt	NNP	<B>	<B>,<B>	-
is	VBZ	Darmstadt	<B>,NNP	-
located	VBN	is	NNP,VBZ	-
in	IN	located	VBZ,VBN	-
GERMANY	NNP	in	VBN,IN	located_in
.	$.	GERMANY	IN,NNP	-

Example: by smaller items it is composed of

DOCUMENT	TF: count $w_1..w_n$					TF*IDF				
	w 1	w 2	w 3		w n	w 1	w 2	w 3		w n
This is a document with text. It contains a number of sentences, just like any other document.	3	0	2		0	0	0	.27		0
Many documents are short, but some are extremely long.	5	0	3	…	1	0	0	.24	…	.21
Principles of Democracy in Bavaria 1. The more you drink, the more the idea of democracy becomes visible	7	4	0		0	0	.62	0		0
DF	**3**	**1**	**2**	…	**1**					

feature vector of the second document

So, more features is better?

- When learning a classifier, every feature introduces a parameter that has to be learned
- The more parameters must be estimated, the more training data we need to do this reliably

➔ might have introduced too many features for the data available

Further problems

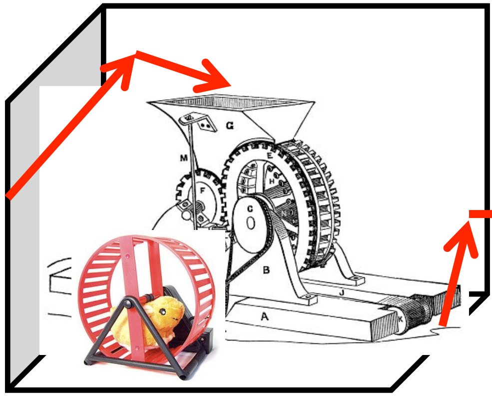
- some ML algorithms assume features to be independent. Dependent features lead to deficient models
- some ML algorithms only deal with binary (numeric, nominal) features
- complex features take long to compute
- sparse features don't help in many cases

➔ need **feature selection** phase to determine the 'good' features

Supervised learning of a classifier

- Classification: assigning (predefined) classes to items
- supervised learning:
  - have a training set with items and their classes
  - train a ML classifier
  - can use the classifier on unseen items

ITEM	f1	f2	f3	…	fn	CLASS
	a	0	.27		T	0
	b	0	.24		F	1
	c	.62	0		T	0
	b	.05	0		F	?



Bayes' Theorem and the Naïve Bayes Classifier

Bayes' Theorem lets us swap the order of dependence between events: We can calculate *P(B|A)* in terms of *P(A|B)*. It follows from the definition of conditional probability and the chain rule that:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

or for *disjoint* C_j forming a **partition** of A:

$$P(C_j \mid A) = \frac{P(A \mid C_j)P(C_j)}{\sum_{i=1}^{n} P(A \mid C_i)P(C_i)}$$

Naïve Bayes Classifier:

- The Naïve Bayes classifier assigns an instance s_k with attribute values (A_1=v_1, A_2=v_2, …, A_m=v_m) to class C_i with maximum P(C_i|(v_1, v_2, …, v_m)) for all *i*.
- The Naïve Bayes classifier exploits the Bayes' s rule and assumes independence of attributes.

Naïve Bayes Classifier

- Likelihood of s_k belonging to C_i

$$= P\Big(C_i \mid \big(v_1, v_2, \ldots, v_m\big)\Big) = \frac{P\Big(\big(v_1, v_2, \ldots, v_m\big) \mid C_i\Big) P(C_i)}{P\Big(\big(v_1, v_2, \ldots, v_m\big)\Big)}$$

- Likelihood of s_k belonging to C_j

$$= P\Big(C_j \mid \big(v_1, v_2, \ldots, v_m\big)\Big) = \frac{P\Big(\big(v_1, v_2, \ldots, v_m\big) \mid C_j\Big) P(C_j)}{P\Big(\big(v_1, v_2, \ldots, v_m\big)\Big)}$$

- Therefore, when comparing P(C_i| (v_1, v_2, …, v_m)) and P(C_j |(v_1, v_2, …, v_m)), we only need to compute P((v_1, v_2, …, v_m)|C_i)P(C_i) and P((v_1, v_2, …, v_m)|C_j)P(C_j)

Naïve Bayes Classifier

- Under the assumption of independent attributes

$$P\left(\left(v_1, v_2, \ldots, v_m\right) \mid C_j\right)$$
$$= P(A_1 = v_1 \mid C_j) \cdot P(A_2 = v_2 \mid C_j) \cdot \ldots \cdot P(A_m = v_m \mid C_j)$$
$$= \prod_{h=1}^{m} P(A_h = v_h \mid C_j)$$

- Furthermore, $P(C_j)$ can be computed by

$$\frac{\text{number of training samples belonging to } C_j}{\text{total number of training samples}}$$

Example: Naïve Bayes Classifier

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

The weather data, with counts and probabilities

outlook			temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

A new day

outlook	temperature	humidity	windy	play
sunny	cool	high	true	?

Likelihood of "yes": $= \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = 0.0053$

Likelihood of "no": $= \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = 0.0206$

➔ Classifier says "no"

Naïve Bayes for Text Classification based on Words as Features

- Multinomial Naïve Bayes Model: The probability of document *d* belonging to class *c* is proportional to the product of the probabilities of terms *t* belonging to class *c*, and to the class prior *P(c)*:

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

- The best class *c* for a document *d* is found by selecting the class, for which the maximum a posteriori (map) probability is maximal:

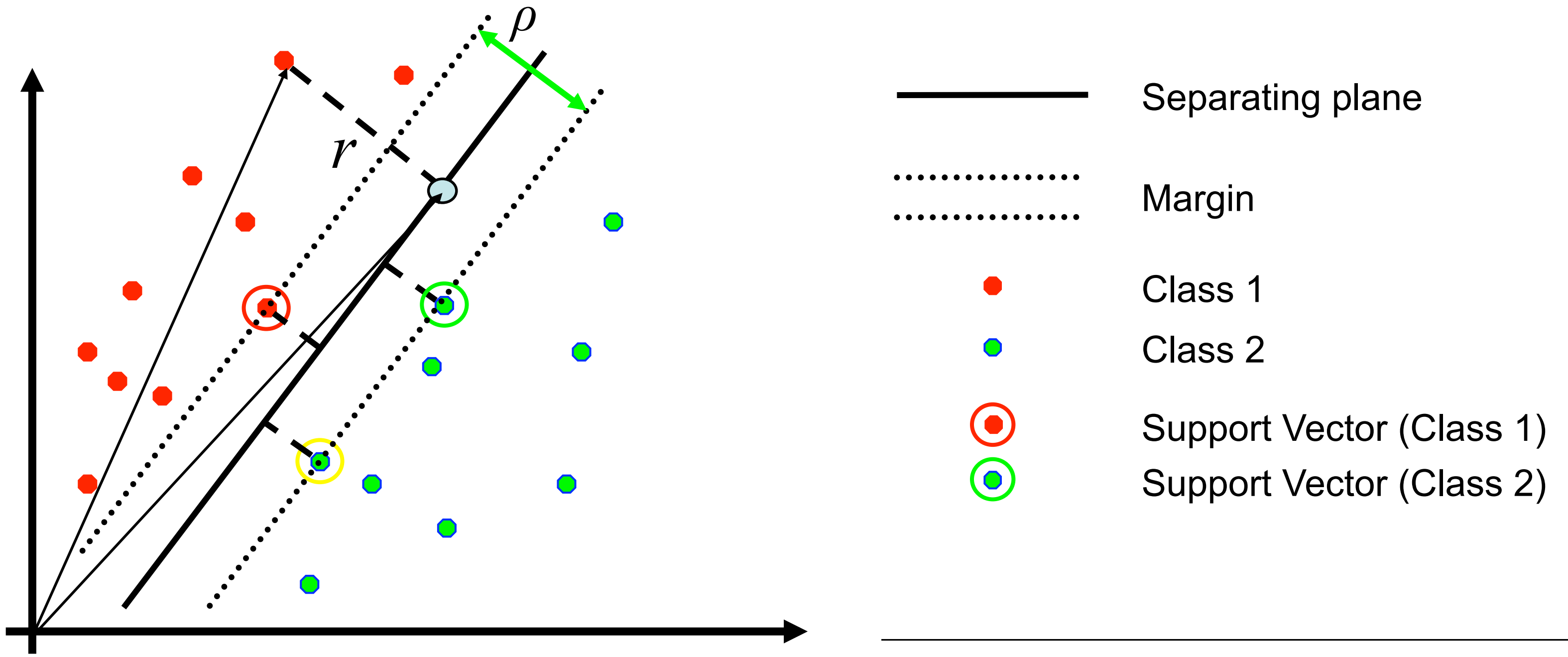
$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} \hat{P}(c|d) = \arg\max_{c \in \mathbb{C}} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c).$$

Summary on Naïve Bayes

- Bayesian methods provide the basis for probabilistic learning methods that use knowledge about the prior probabilities of hypotheses and about the probability of observing data given the hypothesis
- Bayesian methods can be used to determine the most probable hypothesis given the data
- The Naïve Bayes classifier is useful in many practical applications, e.g. text classification
- Training of Naïve Bayes classifiers is very fast
- Binary, numeric and nominal features can be mixed
- Naïve Bayes fails if the independence assumption is violated too much. Especially identical or highly overlapping features pose a problem that has to be addressed with proper feature selection

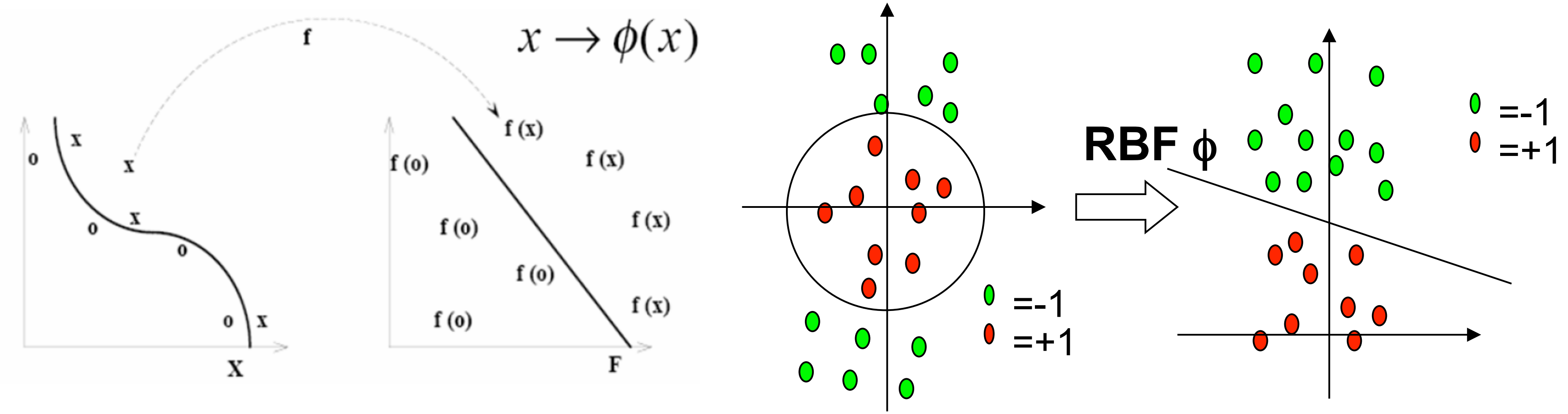
Support Vector Machines

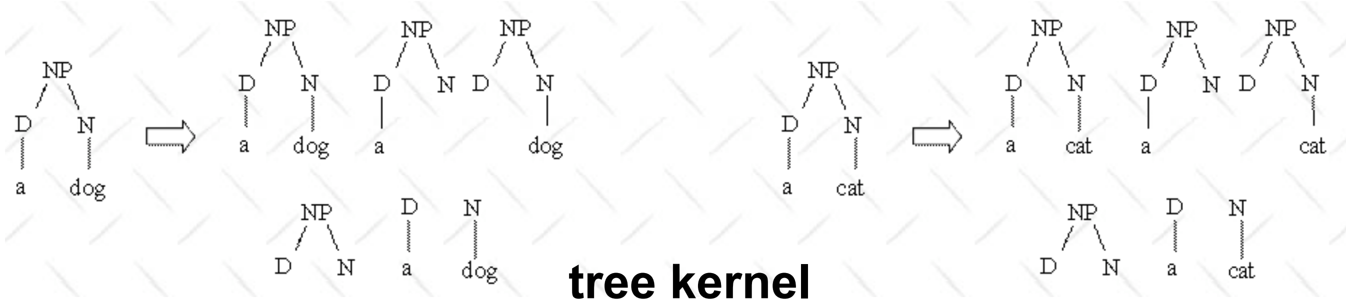
- 2-class classifier, classes +1 and -1; numeric features
- finds the max margin separating hyperplane between two linearly separable classes
- by finding support vectors for classes: Find a separating hyperplane with largest margin
- margin is “soft”: there can be errors that still do not cause the margin to change



Non-linear data: SVM Kernel

- data can be transformed into another (high-dimensional) space
- In this space, data is linearly separable; we train the SVN on this and transform the input data using a **kernel function** Φ(x)





tree kernel

SVM advantages and limitations

Advantages

- deals well with high-dimensional, sparse vectors: can convert nominal features into e.g. Boolean representation
- very flexible: different kernel functions, variation in number of support vectors
- robust classifiers, noise tolerant
- efficient implementations available

Limitations

- only two classes: must use “1 vs. all” or other schemes for multiclass problems
- high computational complexity
- choice of kernel function and its parameters is a trial-and-error enterprise
- black box

Applications of Text Classification:

Applications of text classification in the IR context:

- Content vs. boilerplates for zoning
- Spam detection
- SafeSearch content filtering
- email sorting
- vertical search: scholar, books, shopping, maps, Q&A …

All these classification tasks can be realized by statistical text classification

Sequence Tagging

- We want to know properties of words for further processing, e.g. word classes, names, etc.
- It is possible to learn a method that assigns these properties from **labeled training text**.
- In Machine Learning, this is a classification task. If the sequence of events is taken into account, this is called **sequence tagging**

Examples for tagged text:

- Part-of-Speech:

  `I/`**`PRO`** `saw/`**`V`** `the/`**`DET`** `man/`**`N`** `with/`**`P`** `the/`**`DET`** `saw/`**`N`** `./`**`P`**
- Name tagging:

  `Valerie/`**`B-PERS`** `and/`**`O`** `Rose/`**`B-PERS`** `travel/`**`O`** `to/`**`O`** `New/`**`B-LOC`** `York/`**`I-LOC`** `./`**`O`**

The Sequence tagging problem: Ambiguity, as usual

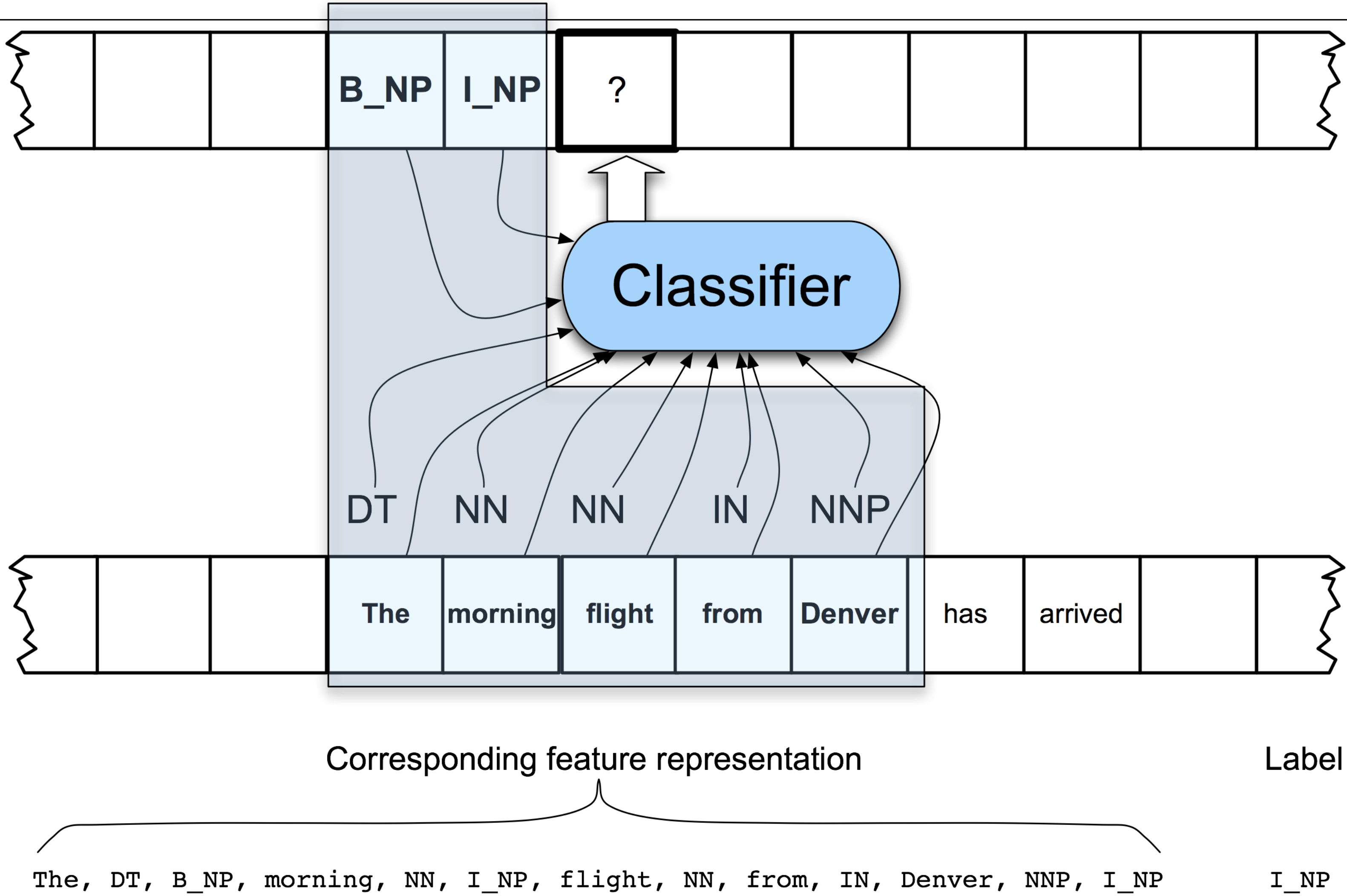
Words often have more than one POS: back

- The *back* door = JJ
- On my *back* = NN
- Win the voters *back* = RB
- Promised to *back* the bill = VB

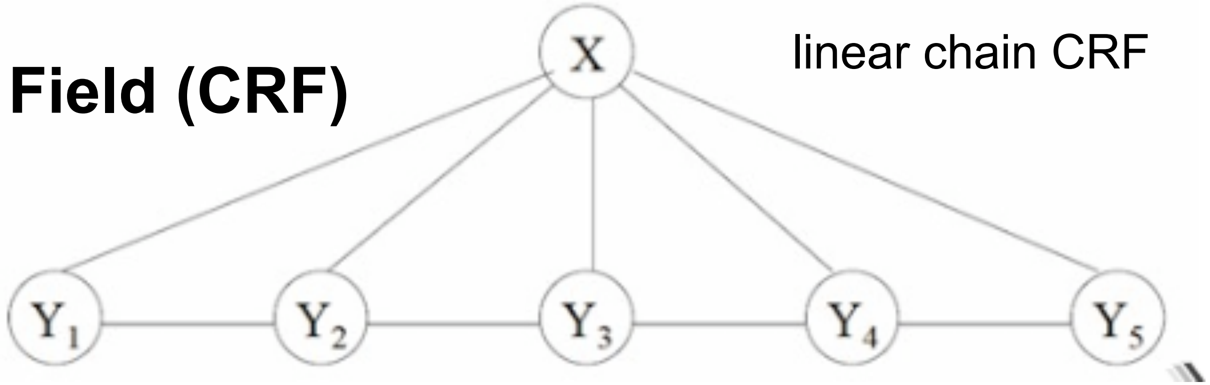
The sequence tagging problem is to determine the label sequence L for a particular sequence of words W:

$$L_{\max} = (l^1_{\max}, l^2_{\max}, \ldots l^T_{\max}) = \underset{L}{\operatorname{argmax}} \; P(L \mid W)$$

Sequence classification with Conditional Markov Models (CMM), e.g. MEMM



Linear Chain
Conditional Random Field (CRF)



linear chain CRF

- Simplest form of CRF: every hidden state has two neighbors
- Markov property in linear chain CRFs: use a variant of Viterbi decoding for computing the optimal label sequence, conditioned on the observed features. This makes efficient decoding possible
- Linear chain CRF subsumes Hidden Markov Model (HMM) but is more expressive, since it allows arbitrary dependencies on the observation sequence

$$P(\mathbf{L} \mid o^1 \ldots o^T) \propto \exp\left(\sum_{e \in E,k} \lambda_k f_k(e, \mathbf{L}|_e, o^1 \ldots o^T) + \sum_{v \in V,k} \mu_k g_k(v, \mathbf{L}|_v, o^1 \ldots o^T) \right)$$

- (see Algorithms of Language Technology for more details)

Properties of CRF

$$L_{\max}^{CRF} = \underset{l^1,\ldots l^T}{\operatorname{argmax}} \frac{1}{Z(o^1..o^T)} \prod_{t=1}^{T} \exp(\mathbf{w}^{\mathbf{T}} \mathbf{f}(l^t, l^{t-1}, o^t)) = \underset{l^1,\ldots l^T}{\operatorname{argmax}} \exp\left(\sum_{t=1}^{T} \sum_{i=1}^{k} \lambda_i f_i(l^t, l^{t-1}, o^t)\right)$$

- Idea: Allow some transitions to vote more strongly than others, depending on the observations
- CRF solves the label bias problem by **normalizing over the whole observation sequence,** unlike the CMM
- like CMM, it is a **discriminative exponential** model
- the marginal probability of the observation sequence is not modeled.
- it is straightforward to implement features on the observation sequence, this includes modeling of dependencies on previous and future observations

Evaluation on POS tagging

model	*error*	*oov error*
HMM	5.69%	45.99%
MEMM	6.37%	54.61%
CRF	5.55%	48.05%
MEMM$^+$	4.81%	26.99%
CRF$^+$	4.27%	23.76%

$^+$Using spelling features

- First order models (bigrams), 45-tagset Penn Treebank, 50% train/test
- With no additional features, HMM and CRF are about equal, and much better than MEMM
- Using additional features, CRF is much better than MEMM.

Summary on Statistical Sequence Tagging

- CMMs are an alternative to HMMs that make it easier to incorporate arbitrary features on the observations
  - discriminative model: can use any classifier, e.g. MaxEnt, SVM etc
  - per-state normalization leads to the label bias problem
- CRF subsumes the advantages of HMM and CMM:
  - per sequence normalization: no label bias problem
  - discriminative: arbitrary features possible
  - downside: slow training

➔especially for small amounts of training data, CRFs are the framework of choice for modern sequence taggers

Next Lecture

Introduction to Information Retrieval

Readings for This Lecture

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. http://nlp.stanford.edu/IR-book/

Mandatory:

- Chapter 3 (pages 49-65): Dictionaries and tolerant retrieval
- Chapter 4.1-4.3 (pages 67-73): Index construction
- Chapter 20 (pages 405-419): Web crawling and indexes