

# Foundations: Statistical Classification in Natural Language Processing

Dietrich Klakow



# What is Classification?



Classification: telling things apart



© Dr. Paul Richardson



---

# Introduction



# Spam/junk/bulk Emails

---

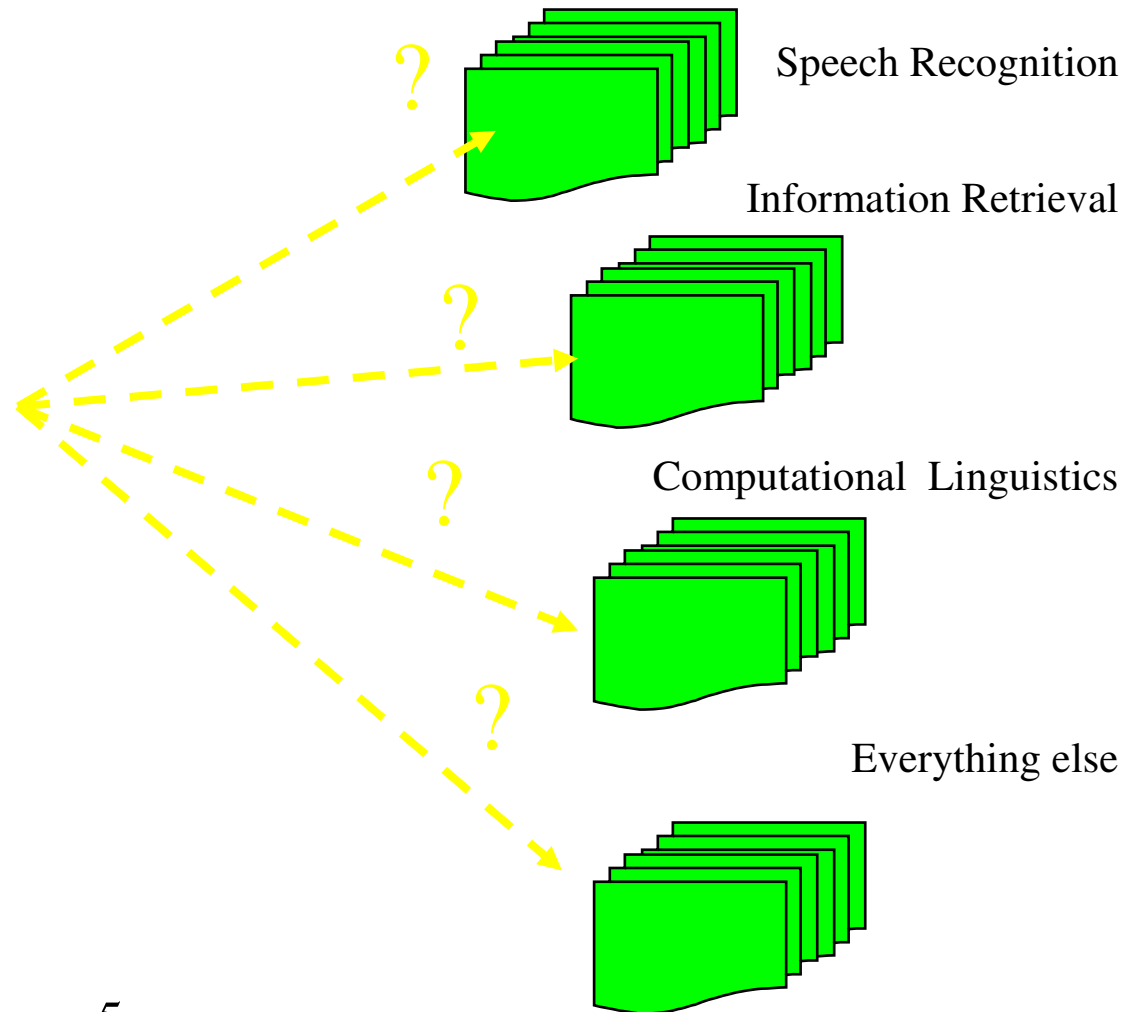
- The messages you spend your time with just to delete them
  - Spam: do not want to get, unsolicited messages
  - Junk: irrelevant to the recipient, unwanted
  - Bulk: mass mailing for business marketing (or fill-up mailbox etc.)

**Classification task: decide for each e-mail whether it is spam/not-spam**



# Text Classification

e.g. text classification





# Question type classification in question answering



Question	Type	Sub-type
Who killed Gandhi ?	HUMAN	individual
Who has won the most Super Bowls ?	HUMAN	group
What city did Duke Ellington live in?	LOCATION	city
Where is the highest point in Japan ?	LOCATION	mountain
What do sailors use to measure time ?	ENTITY	technique
Who is Desmond Tutu ?	DESCRIPTION	human

50 different question types

Most frequent question types:

Human:individual 18%

Location:other 9%

Decription:definition 8%



# Examples of Senses of the Word “Band” from SENSEVAL



band 532732 strip n band/2/1  
band 532733 stripe n band/2/1.2  
band 532734 range n band/2/2  
band 532735 group n band/1/2  
band 532736 mus n band/1/1  
band 532744 brass n brass\_band  
band 532745 radio n band/2/2.1  
band 532746 vb v band/1/3  
band 532747 silver n silver\_band  
band 532756 steel n steel\_band  
band 532765 big n big\_band  
band 532782 dance n dance\_band  
band 532790 elastic n elastic\_band  
band 532806 march n marching\_band

band 532814 man n one-man\_band  
band 532838 rubber n rubber\_band  
band 532903 ed n band/2/3  
band 532949 saw n band\_saw  
band 532963 course n band\_course  
band 532979 pl n band/2/4  
band 533487 vb2 a band/2/5  
band 533495 portion n band/2/1.3  
band 533508 waist n waistband  
band 533520 ring n band/2/1.4  
band 533522 sweat n sweat\_band  
band 533580 wrist n wristband//1  
band 533705 vb3 v band/2/6  
band 533706 vb4 v band/2/7



## Example 1:

---

The incidence of accents and rests, permuted through a regular space-time grid, becomes rhythmic in itself as it modifies, defines and enriches the grouping procedure. For example, a traditional American jazz `<tag` `????` `>band</>` was subdivided into a front line (melodic) section, usually led by trumpet, and rhythm section, usually based on drums.





## Example 1:

---

The incidence of accents and rests, permuted through a regular space-time grid, becomes rhythmic in itself as it modifies, defines and enriches the grouping procedure. For example, a traditional American jazz `<tag "532736">band</>` was subdivided into a front line (melodic) section, usually led by trumpet, and rhythm section, usually based on drums.

`band 532736 mus n band/1/1`



## Example 2:

---

The headsail wardrobe currently consists of a non-overlapping working jib set on a furler, originally designed to cope with wind speeds between 10 and 35 knots plus. But Mary feels it is too small for the lower wind speeds, so she may introduce an overlapping furler for the 10 to 18 knot ???? band.



## Example 2:

---

The headsail wardrobe currently consists of a non-overlapping working jib set on a furler, originally designed to cope with wind speeds between 10 and 35 knots plus. But Mary feels it is too small for the lower wind speeds, so she may introduce an overlapping furler for the 10 to 18 knot `<tag "532734">band</>`.

`band 532734 range n band/2/2`



## Example 3:

---

The Moorsee Lake, on the edge of town, is ideal for swimming. rowing boats are also available for hire. Don't leave without hearing the village brass <tag

???? >band</> which plays three times a week.



## Example 3:

---

The Moorsee Lake, on the edge of town, is ideal for swimming. rowing boats are also available for hire. Don't leave without hearing the village brass `<tag "532744">band</>` which plays three times a week.

`band 532744 brass n brass_band`



## Example 4:

---

Here, suspended from Lewis's person, were pieces of tubing held on by rubber `<tag`  
`????` bands`</>`, an old wooden peg, a bit of cork.



## Example 4:

---

Here, suspended from Lewis's person, were pieces of tubing held on by rubber `<tag "532838">bands</>`, an old wooden peg, a bit of cork.

band 532838 rubber n rubber\_band



# Example for Part-Of-Speech Tagging

---

Xinhua News Agency , Guangzhou , March 16  
( Reporter Chen Ji ) The latest statistics show  
that from January through February this year  
, the export of high-tech products in  
Guangdong Province reached 3.76 billion US  
dollars , up 34.8% over the same period last  
year and accounted for 25.5% of the total  
export in the province .





# Example for Part-Of-Speech Tagging

---

Xinhua/NNP News/NNP Agency/NNP ./,  
Guangzhou/NNP ./, March/NNP 16/CD (/(  
Reporter/NNP Chen/NNP Ji/NNP )/SYM The/DT  
latest/JJS statistics/NNS show/VBP that/IN from/IN  
January/NNP through/IN February/NNP this/DT  
year/NN ./, the/DT export/NN of/IN high-tech/JJ  
products/NNS in/IN Guangdong/NNP  
Province/NNP reached/VBD 3.76/CD billion/CD  
US/PRP dollars/NNS ./, up/IN 34.8%/CD over/IN  
the/DT same/JJ period/NN last/JJ year/NN and/CC  
accounted/VBD for/IN 25.5%/CD of/IN the/DT  
total/JJ export/NN in/IN the/DT province/NN ./.



# Penn-Tree-Bank Tags-Set

- 45 Tags
- Examples:

Tag	Description	Example
CC	Coordinating Conjunction	and, but, or
CD	Cardinal number	one, two, three
DT	Determiner	a. the
JJ	Adjective	yellow
NN	Noun, sing. or mass	province
NNP	Proper noun, singular	IBM
RB	Adverb	quickly, never
VB	Verb, base form	eat
VBD	Verb, past tense	ate
...	...	...



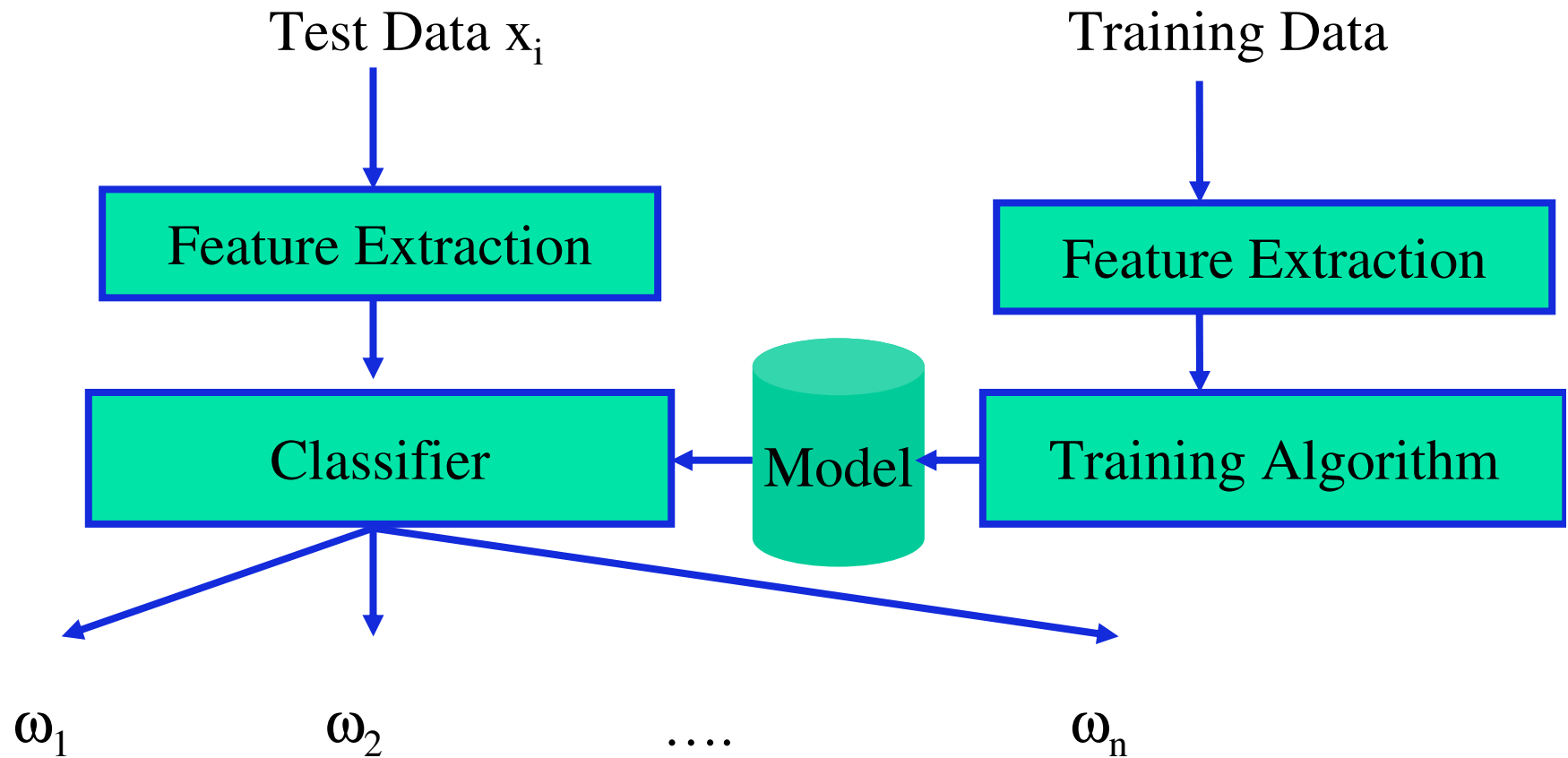
# Definition

---

Pattern Classification:  
Automatic transformation of data  $x_i$   
(observations, features) into a  
set of symbols  $\omega_i$  (classes).



# Flow of Data in Pattern Classification





---

# The Bayes Classifier



# Classifying e-mail for spam/not-spam



- Simple model:
  - No posterior knowledge (i.e. no measurements)
  - Two classes
    - $\omega_1 = \text{“spam”}$
    - $\omega_2 = \text{“not-spam”}$
  - Given:  $P(\omega_1)$  and  $P(\omega_2)$
  - Goal:
    - Minimize the number of mails that get the wrong label

How would you set up a decision rule?



# Classifying Mail

---

spam

Not-spam

$P(\omega_1)$

$P(\omega_2)$

Classify every e-mail as



# Classifying Mail

---

Incorrectly classified



$P(\omega_1)$

$P(\omega_2)$

Classify every e-mail as not-spam





# Classifying Mail



Classify every e-mail as spam

Smaller number of e-mails with wrong label



# Generalization

---

- Minimize number of wrong labels  
↳ pick class with highest probability

Formal notation:

$$\bar{\omega}_i = \arg \max_{\omega_k} P(\omega_k)$$



# Available Measurements $x$

---

- Feature vector  $x$  from measurement
- Probabilities depend on  $x$   $P(\omega_k | x)$
- Definition conditional probability:

$$P(\omega_k | x) = \frac{P(\omega_k, x)}{P(x)}$$



# Bayes Decision Rule: Draft Version

---

- Bayes decision rule

$$\overline{\omega}_i = \arg \max_{\omega_k} P(\omega_k | x)$$

Ugly: usually  $x$  is measured for a given class  $\omega_k$



# Rewrite Bayes Decision Rule

$$\bar{\omega}_i = \arg \max_{\omega_k} P(\omega_k | x)$$

Use definition of cond. probability

$$= \arg \max_{\omega_k} \frac{P(x | \omega_k) P(\omega_k)}{P(x)}$$

$$P(\omega_k | x) = \frac{P(\omega_k, x)}{P(x)}$$
$$= \frac{P(x | \omega_k) P(\omega_k)}{P(x)}$$

$$= \arg \max_{\omega_k} P(x | \omega_k) P(\omega_k)$$

$P(x)$  does not affect decision



# Bayes Decision Rule

---

$$\overline{\omega}_k = \arg \max_{\omega_k} [P(x | \omega_k) P(\omega_k)]$$



# Terminology

---

Prior:  $P(\omega_k)$

Posterior:  $P(\omega_k | x)$



# Naïve Bayes

---

- $x$  is not a single feature, but a bag of features  
e.g. different key-words for your spam-mail detection system
- Assume statistical independence of features

$$P(\{x_1 \dots x_N\} | \omega_k) \approx \prod_{i=1}^N P(x_i | \omega_k)$$





---

Apply Naïve Bayes  
Classifier to Question Type  
Classification



# What are suitable features to classify questions?

---



- Question word?
- Key words?
- Head word?



# Pointwise Mutual Information

---

Definition

$$pMI(x_i, \omega_j) = \frac{N(x_i, \omega_j)}{N} \log \left( \frac{N(x_i, \omega_j)N}{N(x_i)N(\omega_j)} \right)$$

with

$N(x_i, \omega_j)$  : frequency of co - occurrence of  
feature  $x_i$  with class  $\omega_j$

$N(x_i)$  : frequency of feature  $x_i$

$N(\omega_j)$  : frequency of class  $\omega_j$



# Examples

Type	Feature	pMI(x,ω)	N(x,ω)	P(x ω)/P(x)
NUM:count	many	0.015	322	13.7
HUM:ind	Who	0.013	498	4.46
NUM:count	How	0.011	336	6.23
LOC:other	Where	0.011	253	11.22
DESC:manner	How	0.010	274	7.52
LOC:country	country	0.007	120	32.01
NUM:date	When	0.007	124	26.23
DESC:def	is	0.006	284	3.48



# Use Language Models to estimate Probabilities



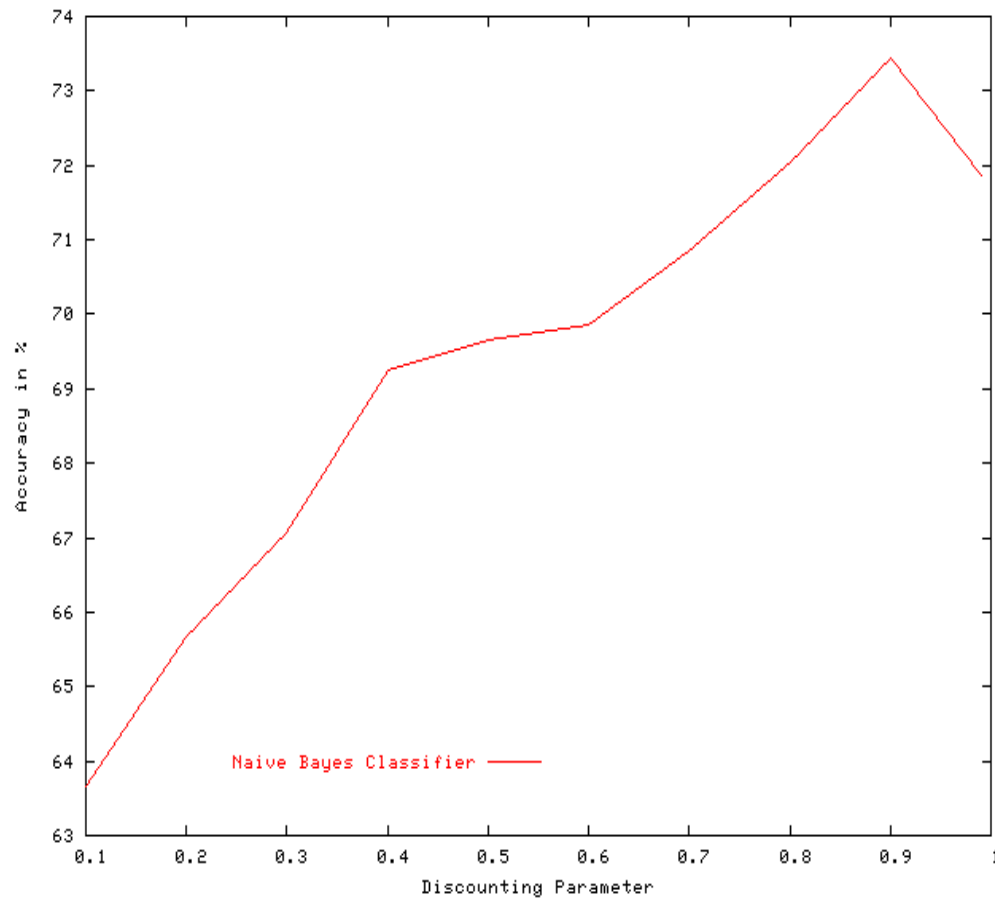
Absolute discounting:

$$P(x_i | \omega_k) = \begin{cases} \frac{N_{\omega_k}(x_i) - d}{N_{\omega_k}} + \alpha \frac{1}{V} & \text{if } N_{\omega_k}(x_i) > 0 \\ \alpha \frac{1}{V} & \text{else} \end{cases}$$

$V$  : size of "feature vocabulary"



# Results



Proper smoothing  
important



---

# How to build a part of speech tagger



# HMM Tagger

---

Specific classification task:

Features: sentence  $W=w_1 \dots w_n$

Class: tag sequence  $T=t_1 \dots t_n$

Bayes classifier:

$$\operatorname{argmax}_T P(W|T)P(T)$$

or

$$\operatorname{argmax}_T P(w_1 \dots w_n | t_1 \dots t_n) P(t_1 \dots t_n)$$





# Simplification of HMM Tagger

---

Assumptions:

word is dependent only on its own POS tag

POS tag depends only on predecessor tag (bigram)

$$\operatorname{argmax}_T [P(w_1|t_1)P(w_2|t_2)\dots P(w_n|t_n)][P(t_1)P(t_2|t_1)\dots P(t_n|t_{n-1})]$$



# Bigram HMM Tagger

---

Estimate

$$P(t_i | t_{i-1}) = N(t_{i-1} t_i) / N(t_{i-1})$$

$$P(w_i | t_i) = N(w_i, t_i) / N(t_i)$$

(or use backing-off-model/absolute discounting)

Compute the most likely sequence using Viterbi algorithm



# Alternative for POS-Tagging: Transformation based learning

---



- Assign each word its most frequent tag ignoring context
- Now apply sequence of transformation rules to correct typical mistakes

“Brill-tagger”



# Alternative Classifiers

---

- Nearest Neighbor
- Support Vector Machines
- Neural Networks
- Decision Trees
- Boosting



# Summary

---



- Many NLP problems can be cast as a classification problem
- Naïve Bayes Classifier often serves as a baseline in statistical NLP