

Introduction to Computational Modeling of  
Lexical and Grammatical Knowledge  
Acquisition using Machine Learning Techniques

December 2004

Damir Čavar

Indiana University

# Agenda

- General Comments
- Probability Theory
- N-Gram Models
  - Frequency, Entropy
- Minimum Description Length Principle
- Vector Space Modeling
- Clustering Algorithms
- Simple Experiments

# Lexical Induction

- Model acquisition of lexical properties
  - If syntactic properties are, or can be described as lexical properties, this also implies modeling of syntactic properties.
  - Cue-based model, where cues are extrinsic and intrinsic properties.
  - Goal: categorization in morpho-syntactic, as well as in semantic or conceptual types.

# Lexical Induction

- Why?
  - The lexicon is the key to language properties.
  - Resolve the paradox: The lexicon is dynamic, language properties are static.
  - Solve some aspects of the Bootstrapping-paradox in language acquisition.
  - Provide some insights and algorithms for lexical acquisition that might have practical relevance for existing computational linguistic problems.

# Modeling Language Acquisition

- The phenomenon refers to:
  - Mapping of non-discrete acoustic events on symbolic representations or activation patterns in a neural net.
  - Segmentation of the symbolic representation, or non-discrete event.
  - Grouping of segments for immediate typing.
  - Grouping of segments for higher level typing.
  - Discovery of relational dependencies for rule induction.

# Lexical Induction

- Instruments
  - Word-, and morphological segmentation
  - Frequency-based methods
  - Minimum Description Length Principle
  - Vector Space Modeling
  - Clustering Analysis
  - Classification

# Introduction

- Jaynes, E.T. (2003) *Probability Theory. The Logic of Science*. Cambridge University Press.
- MacKay, D.J.C. (2003) *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

# Probability Theory

- Plausibility:
  - policeman, night, burglary alarm, jewelry shop, man with mask and bag full of jewels
- Logic deduction based on events vs. Plausibility
- Majority of everyday decisions:
  - Based on incomplete information for deductive reasoning



# Probability Theory

- Plausibility:
  - although we are familiar with plausible conclusions
  - formation of plausible conclusions is a subtle process
  - There is no formal model of this process that is satisfying to everybody working in this domain

# Probability Theory

- Contrast between deductive and plausible reasoning:

- Syllogisms:

- If A is true, then B is true

$$\frac{\text{A is true}}{\text{therefore, B is true}}$$

- inverse:

- If A is true, then B is true

$$\frac{\text{B is false}}{\text{therefore, A is false}}$$

# Probability Theory

- Deductive reasoning along the lines of these syllogisms would be desirable.
- In most situations we do not have the right kind of information for this reasoning:
  - Fallback: weaker syllogisms:
    - If A is true, then B is true

B is true

---

therefore, A becomes more plausible

# Probability Theory

- “Weak” syllogism:
  - The evidence of B being true does not prove that A is true, however
  - verification of one of its consequences does give us more *confidence* in A.
- Weather-example
- Observing B does not give us logical *certainty* that A, but it may induce us to change behavior, plans, *as if we believed* it does.

# Probability Theory

- Another weak syllogism:
  - If A is true, then B is true

A is false

---

therefore, B becomes less plausible

- There is no prove that B is false, but one plausible reason for its being true is eliminated, thus
  - we feel less *confident* about B.
- Scientific reasoning consists usually of the two weak syllogisms.

# Probability Theory

- Another weak syllogism, the policeman reasoning:
  - If A is true, then B becomes more plausible

B is true

---

therefore, A becomes more plausible

- The argument of the policeman is weak.
- Nevertheless, it has a very strong convincing power, almost the power of deductive reasoning.

# Probability Theory

- Cognitive perspective:
  - The brain decides whether something is more or less *plausible*.
  - It evaluates the *degree of plausibility* in some way.
  - It makes use of *old information*.
  - It makes use of the *specific new data* of the problem.
- Reasoning:
  - We depend on *prior information* to help us evaluating the degree of plausibility in a new problem.
  - This is an unconscious process, quite complicated.  
(we call it *common sense*)

# Probability Theory

Probability theory is nothing but common sense reduced to calculation.

Laplace, 1819



# Probability Theory

- Prerequisite: Boolean Algebra
- Representation of degree of plausibility by real numbers.
- Qualitative correspondence with common sense.
- Consistency.

# Probability Theory

- The chance of a particular outcome occurring is determined by the ratio of the number of favorable outcomes to the total number of outcomes.

$$P(A) = \frac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}}$$

- Approach: frequency based

# Probability Theory

- Examples:
  - well-shuffled deck of cards:
    - number of cards 52
  - What is the probability of drawing an ace?

# Probability Theory

- Deck of cards:
  - 4 aces
  - 52 number of cards

$$P(\text{randomly drawing an ace}) = \frac{4}{52} = 0.077$$

- Probability expressed as decimal range between 0 and 1
  - 0 = no chance
  - 1 = certainty

# Probability Theory

- Uniform Distribution:
  - Every outcome has equal likelihood.
- Disjoint outcomes:
  - Outcomes may not occur at the same time.  
(mutually exclusive outcomes)
    - The outcome of drawing just one card can not be an ace and a 9.

# Relative Frequency Theory

- If an experiment is repeated an extremely large number of times and a particular outcome occurs a percentage of the time, then the particular percentage is close to the probability of that outcome.

# Simple Events

- Simultaneously tossing coins:
  - a penny
  - a nickel
  - a dime
- Mutually exclusive events:
  - Head or tail, not both.
- What is the probability of three heads?

# Simple Events

- Total outcomes:

<b>outcome</b>	<b>penny</b>	<b>nickel</b>	<b>dime</b>
1	H	H	H
2	H	H	T
3	H	T	H
4	H	T	T
5	T	H	H
6	T	H	T
7	T	T	H
8	T	T	T



# Simple Events

- Total outcomes: 8
- Favorable outcomes: 1

$$P(3H) = \frac{1}{8} = 0.125$$

- What is the probability of at least two coins landing head?

# Simple Events

- Total outcomes: 8
- Favorable outcomes: 4

$$P(\text{min } 2H) = \frac{4}{8} = 0.5$$

- What is the probability of exactly one coin landing head?

# Simple Events

- Total outcomes: 8
- Favorable outcomes: 3

$$P(1H) = \frac{3}{8} = 0.375$$

# Simple Events

- Independent Events
  - Outcomes that are not affected by other outcomes.
- Dependent Events
  - Outcomes that are affected by other outcomes.
- Dependent Events: Example
  - Randomly drawing an ace from one deck of cards.
  - Randomly drawing another ace from the same deck of cards without returning the first.

# Dependent Events

- 1<sup>st</sup> draw:
  - $P(A) = 4/52 = 0.0769$
- 2<sup>nd</sup> draw:
  - Possibility 1: 1<sup>st</sup> card is not an ace
    - Total number of outcomes: 51
    - Favorable outcomes: 4
    - $P(A) = 4/51 = 0.0784$
  - Possibility 2: 1<sup>st</sup> card is an ace
    - Total number of outcomes: 51
    - Favorable outcomes: 3
    - $P(A) = 3/51 = 0.0588$

# Independent Events

- 1<sup>st</sup> draw:
  - $P(A) = 4/52 = 0.0769$
  - Return card to deck.
- 2<sup>nd</sup> draw:
  - Possibility 1: 1<sup>st</sup> card is not an ace
    - Total number of outcomes: 52
    - Favorable outcomes: 4
    - $P(A) = 4/52 = 0.0769$
  - Possibility 2: 1<sup>st</sup> card is an ace
    - Total number of outcomes: 52
    - Favorable outcomes: 4
    - $P(A) = 4/52 = 0.0769$

# Joint Occurrences

- Tossing three coins as a sequence of events:
  - 1<sup>st</sup> penny
  - 2<sup>nd</sup> nickel
  - 3<sup>rd</sup> dime
- Probabilities for Head:
  - penny:  $\frac{1}{2}$ , nickel:  $\frac{1}{2}$ , dime:  $\frac{1}{2}$
- Multiplication rule:
  - The probability of two or more independent events all occurring is the product of their probabilities.

# Joint Occurrences

- Multiplication of probabilities for head:
  - penny:  $\frac{1}{2}$ , nickel:  $\frac{1}{2}$ , dime:  $\frac{1}{2}$
  - $0.5 \times 0.5 \times 0.5 = 0.125$
  - Compare with classical approach!
- Notation:

$$P(AB) = P(A) \times P(B)$$



# Joint Occurrences

- Drawing two aces from one deck of cards without returning the first:

$$P(AB) = \frac{4}{52} \times \frac{3}{51} = \frac{12}{2652} = 0.0045$$

- Drawing two aces from one deck of cards returning the first:

$$P(AB) = \frac{4}{52} \times \frac{4}{52} = \frac{16}{2704} = 0.0059$$

# Mutually Exclusive Events

- Tossing one coin once:
  - What is the probability of at least one outcome head or one outcome tail?

# Mutually Exclusive Events

- Tossing one coin once:
  - The probability of at least one outcome head or one outcome tail is: 1
- Reason:
  - $P(\textit{Head})=0.5$
  - $P(\textit{Tail})=0.5$
  - $P(\textit{Head or Tail})=0.5+0.5=1$
- What is the probability of drawing a king or an ace?

# Mutually Exclusive Events

- The probability of drawing a king or an ace:
  - king:  $4/52$
  - ace:  $4/52$
  - $P(\textit{king or ace}) = 4/52 + 4/52 = 8/52 = 0.154$
- Addition rule
  - Only with mutually exclusive outcomes the probability of one outcome or another outcome is the sum of the probabilities of single outcomes.

# Non-Mutually Exclusive Events

- What is the probability of the outcome of at least one head with one coin tossed twice?
  - head1:  $1/2$
  - head2:  $1/2$
  - $P(\text{min}1H) = 1/2 + 1/2 = 1$
- **No!**
- **No addition rule with non-mutually exclusive events!**

# Non-Mutually Exclusive Events

- The probability of the outcome of at least one head with one coin tossed twice:
  - Notation: At least one favorable outcome in two events

$$P(A + B) = P(A) + P(B) - P(AB)$$

- Read as: probability of A plus the probability of B, minus the joint probability of an occurrence of A and B.
- Why?

# Non-Mutually Exclusive Events

- Total outcomes:

<b>outcome</b>	<b>coin</b>	
1	H	H
2	H	T
3	T	H
4	T	T

- Favorable outcomes: 3
- Addition rule: adds two favorable outcomes from the first toss, and two favorable outcomes from the second toss!

# Conditional Probability

- Data:

<b>Students</b>	<b>Younger than 25</b>	<b>25 or older</b>	<b>sum</b>
Male	20	40	60
Female	5	35	40
sum	25	75	100

- What is the probability that a student selected at random will be male?



# Conditional Probability

- The probability that a student selected at random will be male:
  - $P(\text{male}) = 40/100 = 0.4$
- What is the probability that a person younger than 25 selected at random will be male?

# Conditional Probability

- The probability that a person younger than 25 selected at random will be male:
  - A = being male
  - B = being younger than 25
  - A is conditional upon B
  - 5 of 25 are male and younger than 25

$$P(A|B) = \frac{AB}{B} = \frac{P(AB)}{P(B)}$$

- Note: Reverse set via  $P(B|A)$ !

# Conditional Probability

- The probability that a person younger than 25 selected at random will be male:
  - A = being male
  - B = being younger than 25
  - A is conditional upon B
  - 20 of 25 are male and younger than 25

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

# Conditional Probability

- Conditional probability = posterior probability
  - $P(a|b)$ , given  $a$  and  $b$  as any propositions
  - “the probability of  $a$ , given that  $b$  occurred”
  - “the probability of  $a$ , given that all we know is  $b$ ”

# Conditional Probability

- Definition:

- Conditional probabilities in terms of unconditional probabilities.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- whenever  $P(B) > 0$

# Conditional Probability

- Definition as the Product Rule:
  - Conditional probabilities in terms of unconditional probabilities.

$$P(A \cap B) = P(A|B)P(B)$$

– or . . .

# Conditional Probability

- Definition as the Product Rule:
  - Conditional probabilities in terms of unconditional probabilities.

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$$

- Why?

# Conditional Probability

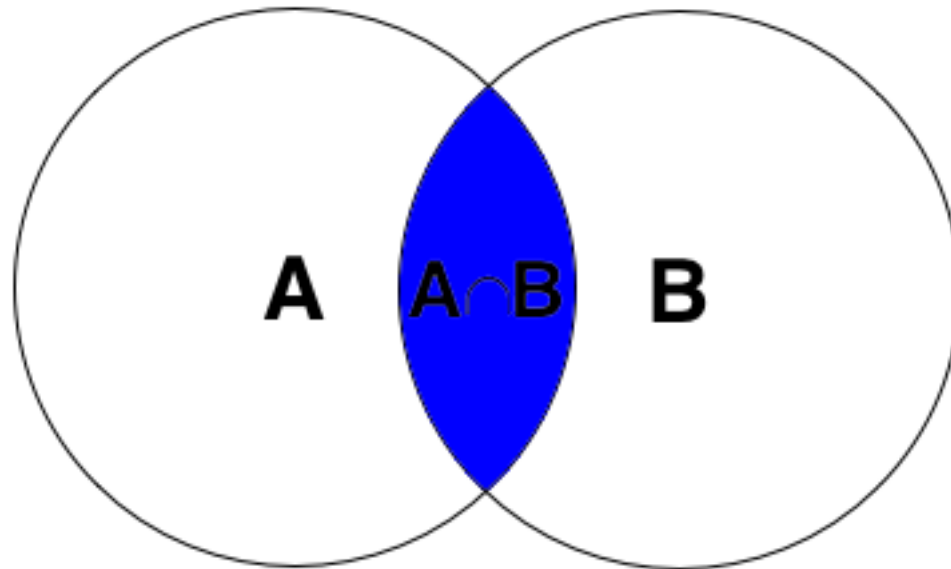
- Product Rule:

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$$

- For  $A$  and  $B$  to be true, we need  $B$  to be true, and we also need  $A$  to be true given  $B$ .
- Commutativity of conjunction!
- Set intersection is symmetric:  $A \cap B = B \cap A$



# Conditional Probability



# Conditional Probability

- Equating the two right-hand sides of the product rule:

$$P(B|A)P(A) = P(A|B)P(B)$$

$$\frac{P(B|A)P(A)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

# Bayes' Theorem

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

- Bayes' theorem, Bayes' law, Bayes' rule
  - Underlies modern AI systems for probabilistic inference.
  - What is it good for?

# Bayes' Theorem

- Properties:
  - Requires three terms (1 conditional & 2 unconditional probabilities) to calculate one conditional probability.
- Use:
  - When we have good probability estimates for these three numbers we can compute the fourth.

# Bayes' Theorem

- Example:
  - 2% of a population has a disease
  - A disease test says that 3.2% of the population has this disease.
  - Chance for testing a person with this disease positive is 75%.
  - What is the probability that a person who is tested positive really has the disease?

# Bayes' Theorem

- Example:
  - $D+ / D-$  is the event of having/not having the disease
  - $T+ / T-$  is the event of a positive/negative test
  - $P(D+) = 0.02$
  - $P(T+) = 0.032$
  - $P(T+ | D+) = 0.75$
- What is  $P(D+ | T+)$  ?

# Bayes' Theorem

	$D+$	$D-$	total
$T+$			0.032
$T-$			0.968
total	0.02	0.98	1.000

# Bayes' Theorem

	$D+$	$D-$	total
$T+$			0.032
$T-$			0.968
total	0.02	0.98	1.000

$$P(D+ | T+) = \frac{P(T+ | D+)P(D+)}{P(T+)}$$



# Bayes' Theorem

	$D+$	$D-$	total
$T+$			0.032
$T-$			0.968
total	0.02	0.98	1.000

$$P(D+ | T+) = \frac{P(T+ | D+)P(D+)}{P(T+)} = \frac{0.75 \times 0.02}{0.032} = 0.46875$$

# Bayes' Theorem

	$D+$	$D-$	total
$T+$	0.015		0.032
$T-$			0.968
total	0.02	0.98	1.000

$$P(D+ | T+) = \frac{P(T+ | D+)P(D+)}{P(T+)} = \frac{0.75 \times 0.02}{0.032} = 0.46875$$

# Bayes' Theorem

	$D+$	$D-$	total
$T+$	0.015	0.017	0.032
$T-$			0.968
total	0.02	0.98	1.000

$$P(D+ | T+) = \frac{P(T+ | D+)P(D+)}{P(T+)} = \frac{0.75 \times 0.02}{0.032} = 0.46875$$

# Bayes' Theorem

	$D+$	$D-$	total
$T+$	0.015	0.017	0.032
$T-$	0.005		0.968
total	0.02	0.98	1.000

$$P(D+ | T+) = \frac{P(T+ | D+)P(D+)}{P(T+)} = \frac{0.75 \times 0.02}{0.032} = 0.46875$$

# Bayes' Theorem

	$D+$	$D-$	total
$T+$	0.015	0.017	0.032
$T-$	0.005	0.963	0.968
total	0.02	0.98	1.000

$$P(D+ | T+) = \frac{P(T+ | D+)P(D+)}{P(T+)} = \frac{0.75 \times 0.02}{0.032} = 0.46875$$

# Probability Distributions

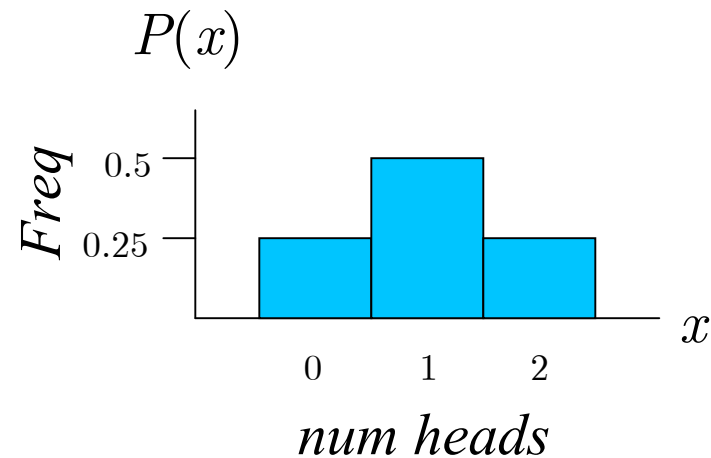
- Pictorial Display of the probability  $P(x)$  for any value of  $x$ .
- Two tossed coins:

<b>outcome</b>	<b>Coins</b>		<b>Num heads</b>
1	H	H	2
2	H	T	1
3	T	H	1
4	T	T	0

# Probability Distributions

- Pictorial Display of the probability  $P(x)$  for any value of  $x$ .
- Two tossed coins:

$x$	$P(x)$
0	$\frac{1}{4}$
1	$\frac{1}{2}$
2	$\frac{1}{4}$



# Probability Distributions

- Use of probability to describe events includes the notion of uncertainty. This can be described with a probability distribution:

- fair coin:

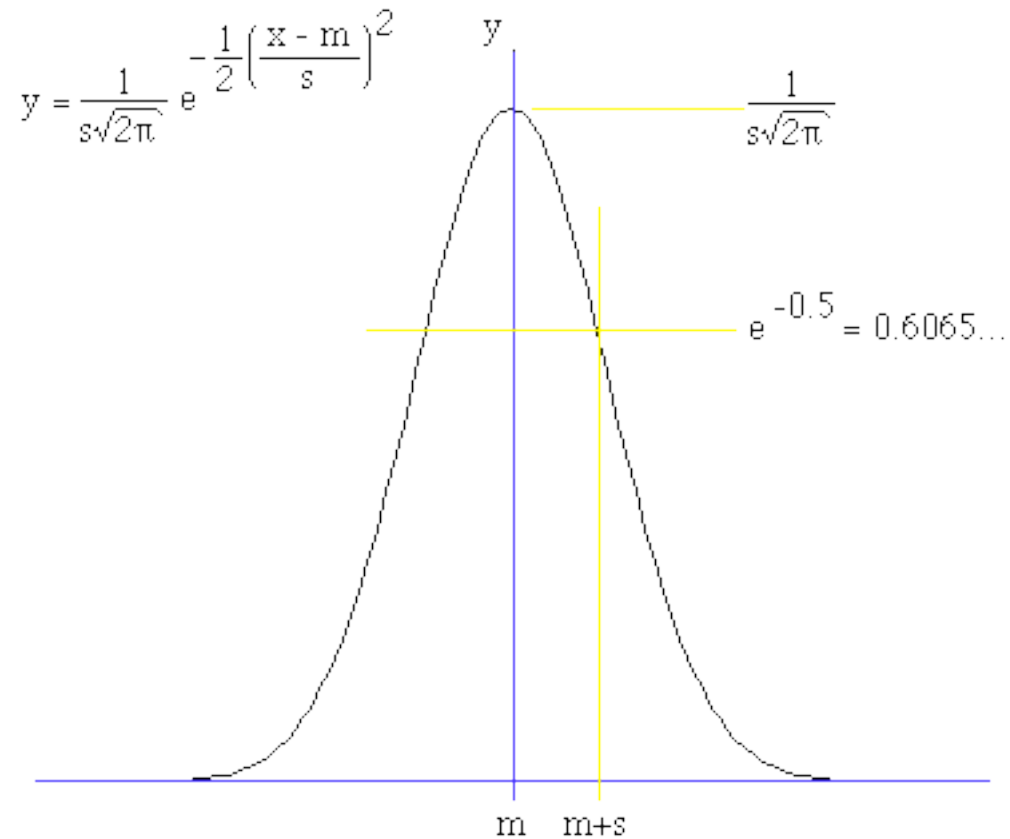
$x$	$P(head)$	$P(tail)$
0	$\frac{1}{2}$	$\frac{1}{2}$

- biased coin:

$x$	$P(head)$	$P(tail)$
0	$\frac{3}{4}$	$\frac{1}{4}$



# Gaussian Normal Distribution



# Uncertainty

- The probability distributions will differ, some coins are more biased than others.
  - We are more uncertain about the outcome of the fair coin than of the biased.
  - How to quantify this notion of uncertainty?
    - Is there a mathematical method to calculate the uncertainty given a probability distribution?
  - Function:
    - Parameter: a probability distribution for a random variable  $X$ 
      - e.g. with  $N$  possible values  $X$  can have,
      - $X = \{ P(n_1), P(n_2), \dots P(n_N) \}$

# Uncertainty

- Properties of the uncertainty function,  $H$ :
  - It returns real values.
  - It should be maximized for the uniform distribution, i.e. this is equivalent to complete uncertainty.
    - Everything is equal likely to occur.
  - It is continuous, i.e. for arbitrary small changes in the probabilities we expect arbitrary small changes in the real value returned.
  - It does not depend on the order or grouping of events, just on the distribution as such.

# Uncertainty

- Maximization requirement:

$H(P(n_1), P(n_2), P(n_3), \dots, P(n_N))$  is max when :  $\forall n : P(n) = \frac{1}{N}$

# Uncertainty

- Independence of Partitioning or Grouping:

$$X = \{P(a) = .5, P(b) = .2, P(c) = .3\}$$

- Outcome of  $b$  or  $c$  occurs 50% of the time:

$$X = \{P(a) = .5, P(Y) = .5\}$$

$$Y = \{P(b) = .4, P(c) = .6\}$$

# Uncertainty

- Entropy (average information content):

$$H[X] = k \sum_{x \in X} P(x) \log P(x)$$

–  $\log_2$  for bits, -1 for positive values,  $0 \log 0 = 0$ :

$$H[X] = - \sum_{x \in X} P(x) \log_2 P(x)$$

$$H[X] = \sum_{x \in X} P(x) \log_2 \frac{1}{P(x)}$$

# N-gram Models

- List all possible symbol combinations of length  $n$  for a given corpus,
  - symbols: phones, phonemes, characters, morphemes, words (tokens or types), sentences, paragraphs etc.
- together with their frequencies (absolute + number of all elements/tokens; relative)

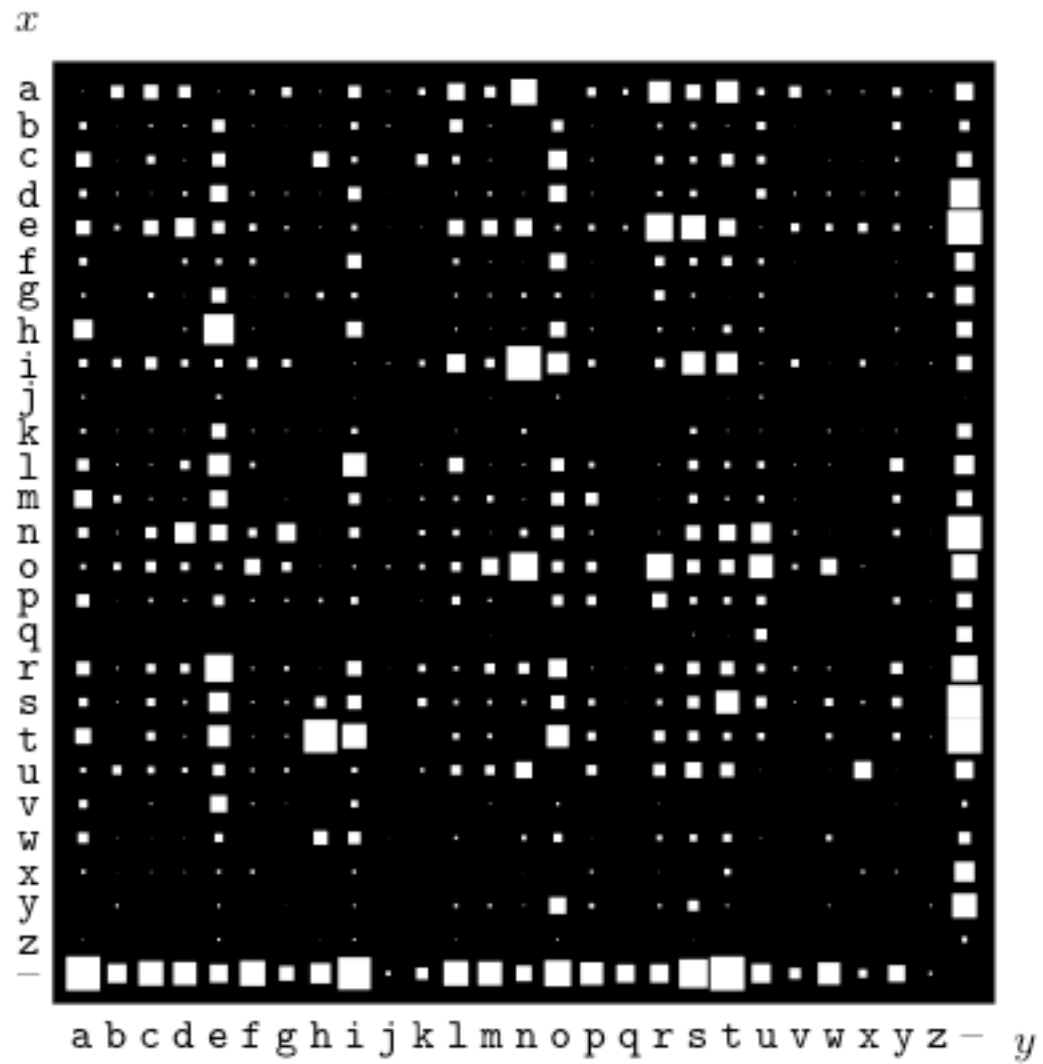
# Frequency Profiles

- Unigram
- Bi-gram
  - Tables/graphics taken from MacKay (2003)



$i$	$a_i$	$p_i$	
1	a	0.0575	a
2	b	0.0128	b
3	c	0.0263	c
4	d	0.0285	d
5	e	0.0913	e
6	f	0.0173	f
7	g	0.0133	g
8	h	0.0313	h
9	i	0.0599	i
10	j	0.0006	j
11	k	0.0084	k
12	l	0.0335	l
13	m	0.0235	m
14	n	0.0596	n
15	o	0.0689	o
16	p	0.0192	p
17	q	0.0008	q
18	r	0.0508	r
19	s	0.0567	s
20	t	0.0706	t
21	u	0.0334	u
22	v	0.0069	v
23	w	0.0119	w
24	x	0.0073	x
25	y	0.0164	y
26	z	0.0007	z
27	-	0.1928	-





# N-gram Scripts

- Wort n-grams
  - frequency.py
  - frequency2.py
  - frequencyNFW.py
  - ngram.py
  - ngramchar.py
  - unigramchar.py

# N-gram Model LID

- Language identification via distributional similarity of n-grams
  - Train language model:
    - extract 3-grams of characters from text for each language, together with the relative frequency of each 3-gram
  - Identify language:
    - extract 3-grams of characters from text
    - compare the standard deviation for each 3-gram with each language model
    - minimum standard deviation identifies the corresponding language

# Information Theory

- Mutual Information

$$I(X;Y) = P(XY) \log_2 \frac{P(XY)}{P(X)P(Y)}$$

- How many bits can we spare by storing  $\langle xy \rangle$  together, rather than each separate?
- How much do we expect  $y$  given  $x$ ?

# Information Theory

- Relative Entropy

$$D(y||x) = p(y) \lg \frac{p(y)}{p(y|x)}$$

- Distance between two distributions:
  - Independent:  $P(y)$
  - Conditional:  $P(y|x)$
- How many bits more would we need to represent  $\langle xy \rangle$  when we store them together, or when we store them as separate units?