Chapter 2: Natural Language as a Sequence of Symbols

What is a Language?

Language?

- Natural language

  E.g. English, German, Urdu, …

- Formal languages

  E.g. C++, Java, LaTeX, …

- Descriptive Languages

  E.g. chemical formulas, DNA

Language

Definition Words and Vocabulary:

The Vocabulary is a set V with W different Symbols w_i.

Remark: we will always call the symbols "words" even though they may be letters, amino acids, ...

Language

Definition Text:

A sequence of symbols from the set V

Definition Language:

Set of all possible texts

Examples of Vocabularies

English (standard):

the

and

or

you

why

..

In some languages a word is not clearly defined

Examples of Vocabularies

Snippet of Chinese text:

风暴造成的主要影响是令墨西哥东南部普降暴雨

No white space segmentation

Examples of Vocabularies: use characters as elementary symbols

Chinese

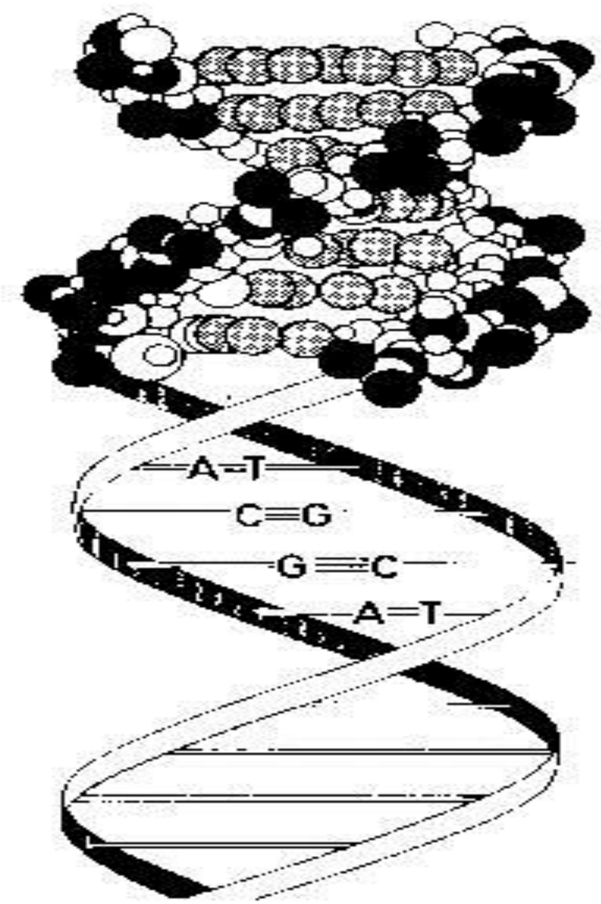
的	不	用	一	新	时	吗	所	被	说	男	事
人	在	之	二	市	也	家	后	只	着	女	得
上	于	要	三	与	还	可	分	都	位	几	真
中	前	好	四	内	出	件	种	做	把	各	对
下	者	了	五	本	去	里	将	已	吧	谁	看
大	会	年	六	地	到	最	很	长	难	找	见
小	号	月	十	这	他	回	而	行	来	子	加
是	我	日	个	此	性	万	数	等	站	字	更
没	和	为	次	建	就	能	天	再	每	那	多
有	你	名	元	全	部	爱	无	以	起	哪	少

English

A, B, C, D, E

..

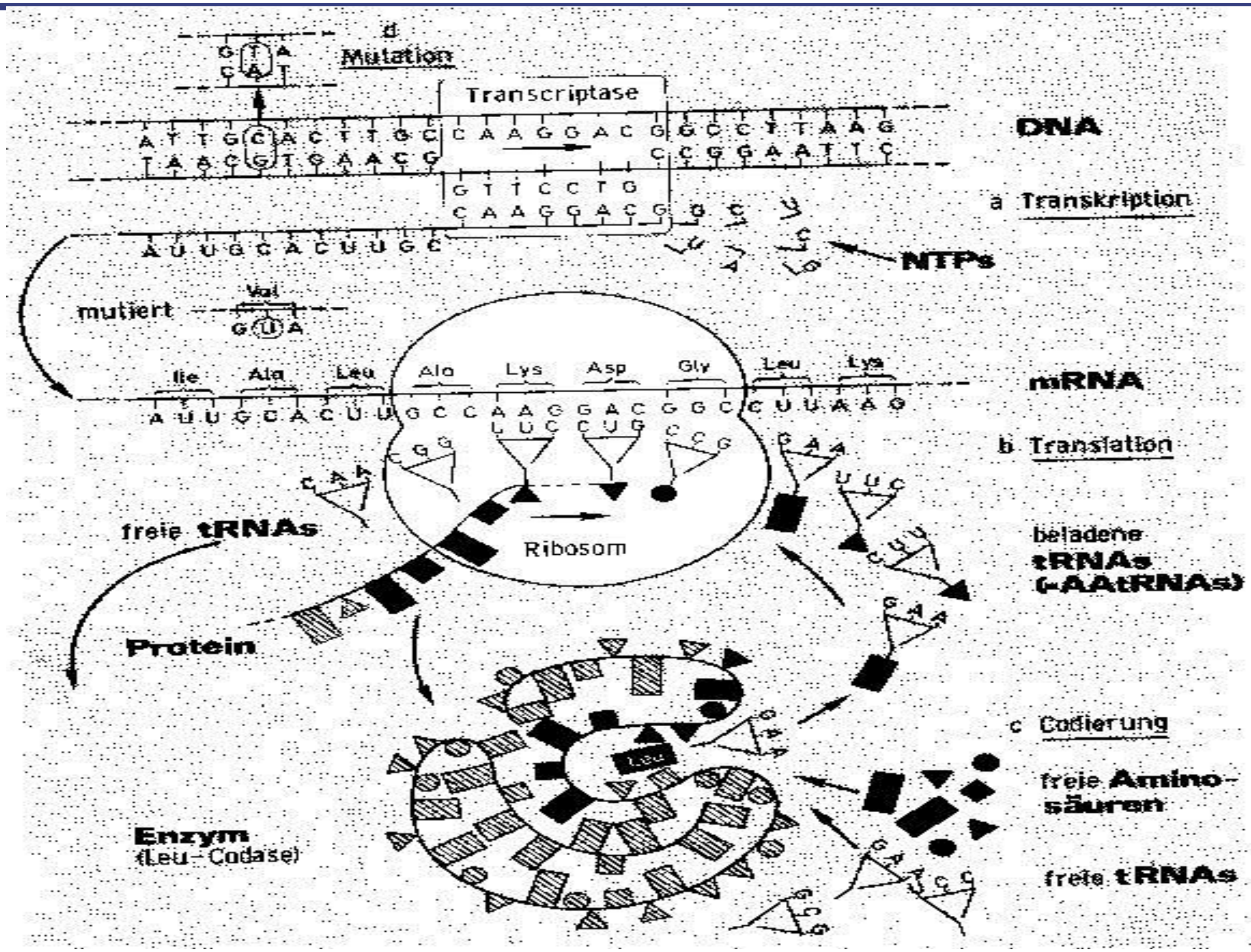
DNA



DNA is a “text” composed of four “symbols/characters” encoding “life”:

Adenin (A)
Guanin (G)
Thymin (T)
Cytosin “(C)

The Use of DNA: encode protein production



Section 2.1. Zipf's Law

See Manning & Schütze section 1.4.3

Motivation

Is natural language (like English) governed by rules (grammar) or statistics?

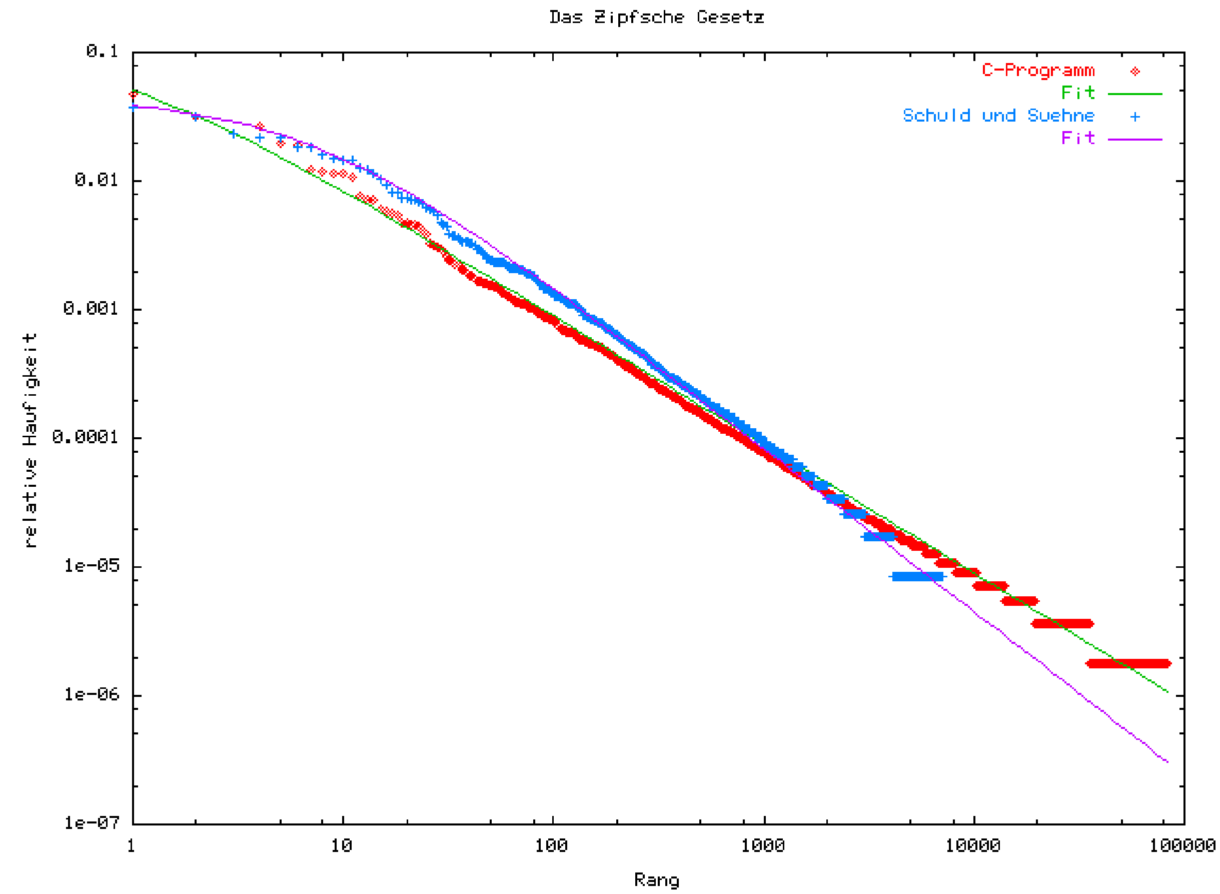
Zipf’s Analysis

- Count the frequency of all the words in a corpus
- Sort the words by frequency
- Rank: position of a word in the sorted list
- Plot rank vs. frequency

Example: “Crime and Punishment” by Dostojewskij

Rank	Word	Frequency
1	THE	4434
2	AND	3746
3	HE	2709
4	TO	2562
5	A	2500
6	DOUBLE-QUOTE	2128
7	END-QUOTE	2118
8	OF	1903
9	IN	1724
10	YOU	1667
11	I	1666
12	IT	1483
13	WAS	1442
□ …	□ ….	□ …

Zipf's Law: two Examples



Zipf's Law

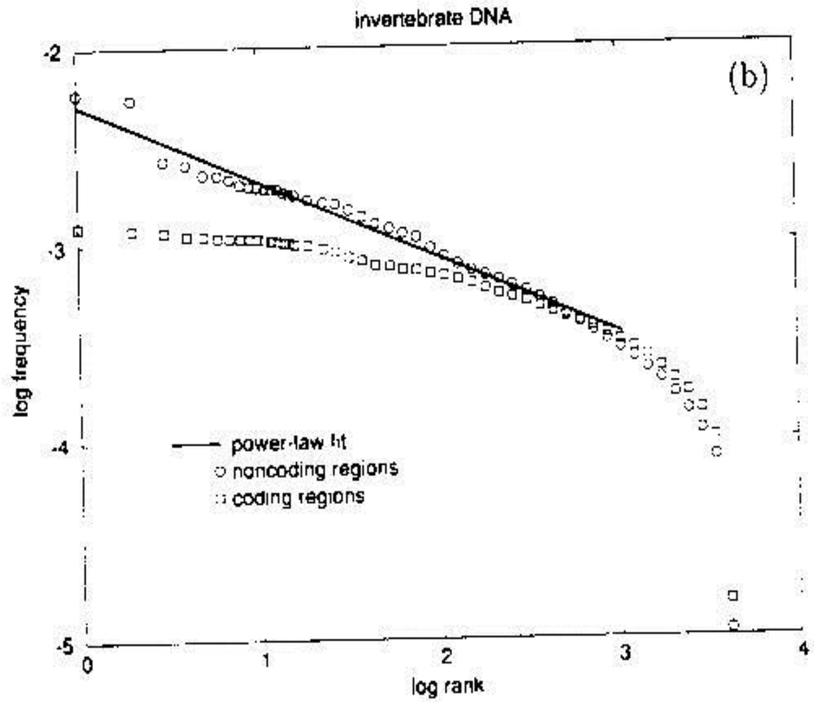
- Mandelbrot Distribution:

$f(r) = m/(c+r)^B$

	C-Program	Crime and Punishment
μ	0.09	0.54
c	0.8	7.0
B	1.0	1.24

Zipf's Law for DNA

- Use a subsequence of fixed length as a word
- DNA also satisfies Zipf's law

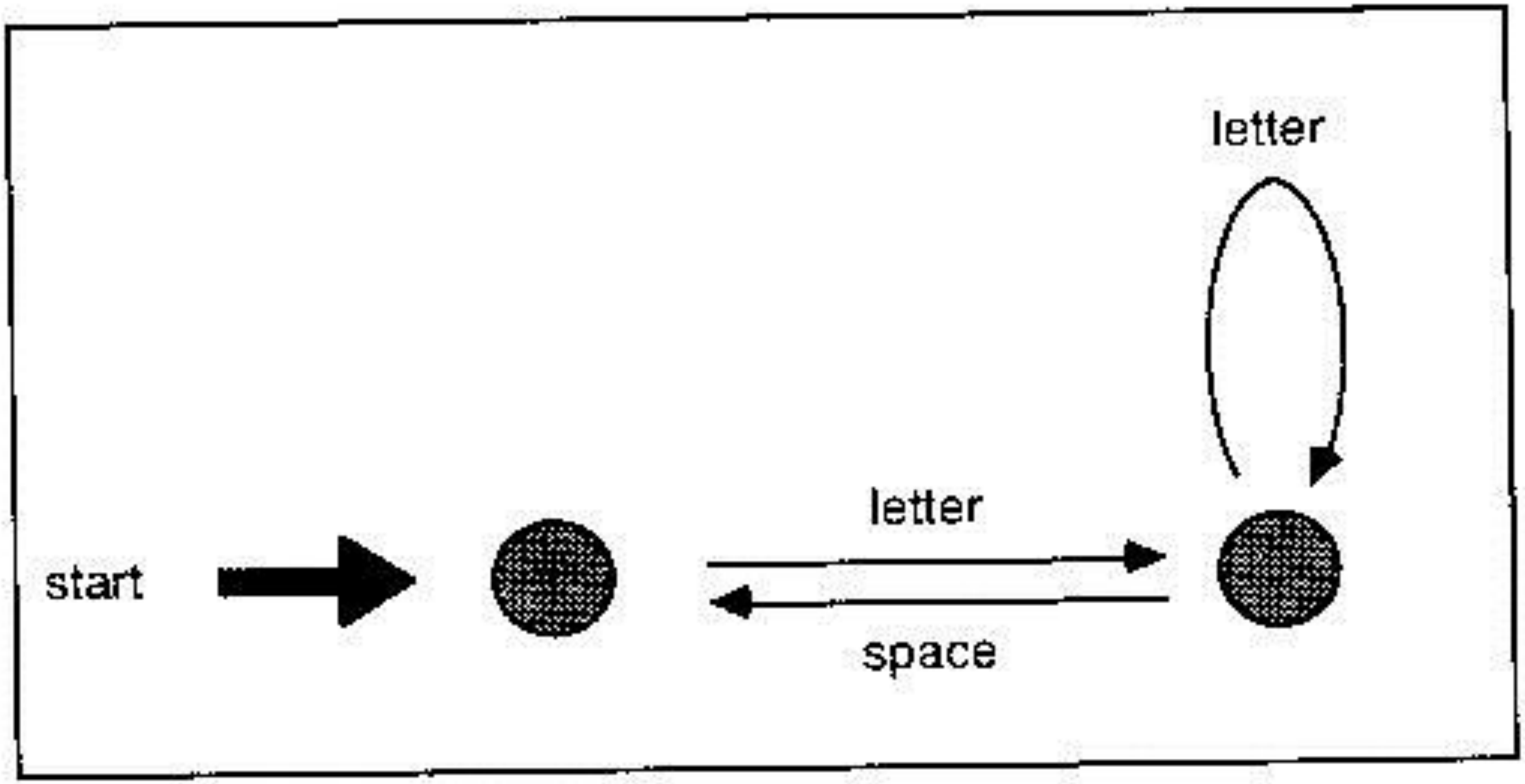


Statistical properties of DNA sequences
C.-K. Peng a,b, S.V. Buldyrev b, A.L. Goldberger a'c, S. Havlin b'd, R.N. Mantegna b,e, M. Simons a, H.E. Stanley b

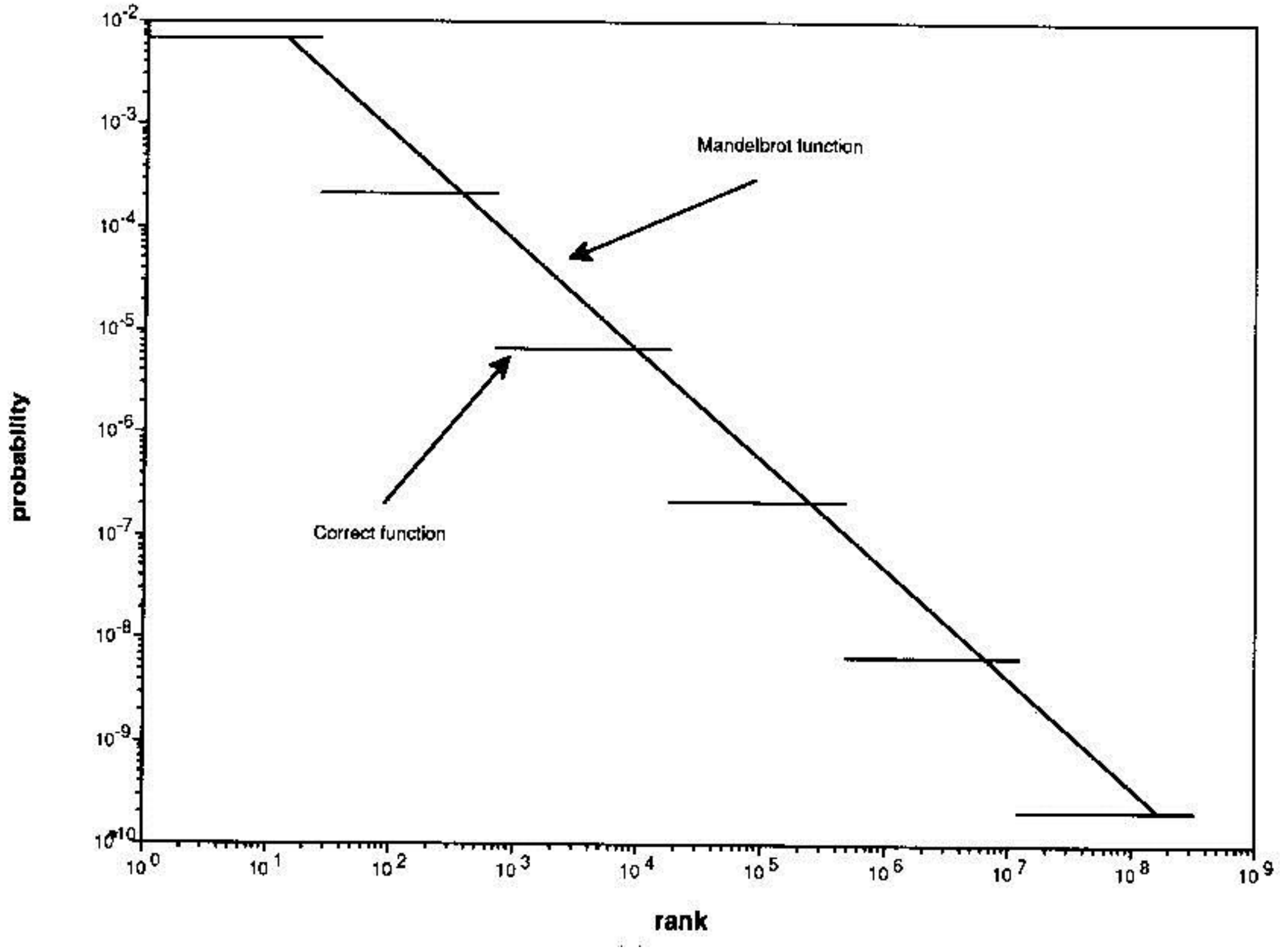
“Derivation” of Zipf’s Law by Miller

- Press a key on the keyboard at random
- “Space” is pressed with probability p
- After “space”, a proper letter has to follow
- All letters except “space” have the same probability

Derivation of Zipf's Law by Miller



Derivation of Zipf's Law by Miller



Section 2.2. Basics of Probability Theory

See Manning & Schütze section 2.1

Goal of this section

- Define some terms and definitions we need

- This is not a tutorial in probability theory!

Definition: Probability of a Word

Positive

$$P(w = w_i) \geq 0 \qquad \forall w_i \in W$$

Normalized

$$\sum_{w_i \in W} P(w_i) = 1$$

Additive

$$P(w = w_i \vee w = w_j) = P(w = w_i) + P(w = w_j)$$

$$\forall w_i \neq w_j$$

Probability of a Sequence of Words

$$P(\text{"to be or not to be"}) =$$

$$\mathrm{P}(\mathrm{w}_1 = \text{"}\textit{to}\text{"}, w_2 = \text{"}\textit{be}\text{"} \ldots\ldots w_6 = \text{"}\textit{be}\text{"})$$

Shorthand notation:

If we have a specific sequence $w_1, w_2, w_3, \ldots w_N$

We denote the probability of this specific sequence by

$$\mathrm{P}(\mathrm{w}_1, w_2, w_3, \ldots\ldots w_N)$$

Left/right Marginal Distribution

Right marginal distribution

$$\sum_{w_1 \in V} \mathrm{P}(\mathrm{w}_1, w_2) = \mathrm{P}(\mathrm{w}_2)$$

Left marginal distribution

$$\sum_{w_2 \in V} \mathrm{P}(\mathrm{w}_1, w_2) = \mathrm{P}(\mathrm{w}_1)$$

What about the general case: $\mathrm{P}(\mathrm{w}_1, w_2, w_3, \ldots.. w_N)$

Expectation value

- Let $f(w_i)$ be some observable
- Expectation value:

$$E[f(V)] = \sum_{w_i \in V} p(w_i) f(w_i)$$

- Example:
  - $f(w_i)$ is the outcome of rolling a dice
  - $E[f(w_i)]$ is the average over many rolls

Expectation value: Example 2

Let:

$f(w_i) = -\log(p(w_i)\ )$

Hence the expectation value is

$$E[-\log(p(V))] = -\sum_{w_i \in V} p(w_i) \log(p(w_i))$$

$E[-\log(p(w_i))]$ is called “entropy” (denoted by H)

Conditional Probability

Definition of conditional probability

$$P(w_2 \mid w_1) = \frac{P(w_1, w_2)}{P(w_1)}$$

Interpretation:
$P(w_2|w_1)$ is the probability that w_2 is observed given that the predecessor word is w_1

Conditional Probability (II)

Bayes theorem:

$$P(w_2 \mid w_1)P(w_1) = P(w_1 \mid w_2)P(w_2)$$

Proof:

$$P(w_2 \mid w_1)P(w_1) = \frac{P(w_1, w_2)}{P(w_1)} P(w_1)$$

$$= P(w_1, w_2)$$

$$= \frac{P(w_1, w_2)}{P(w_2)} P(w_2) = P(w_1 \mid w_2)P(w_2)$$

Statistical Independence for a sequence of 2 words

Definition:

$$P(w_1 \mid w_2) = P(w_1)$$

Consequence:

$$P(w_1, w_2) = P(w_1)P(w_2)$$

Bayes Decomposition: write joint probability as product of conditional probabilities

$$P(w_1, w_2, w_3, \ldots\ldots w_N) =$$

$$= P(w_N \mid w_1, w_2, w_3, \ldots\ldots w_{N-1}) P(w_1, w_2, w_3, \ldots\ldots w_{N-1})$$

$$= P(w_N \mid w_1, w_2, \ldots w_{N-1}) P(w_{N-1} \mid w_1, w_2, \ldots. w_{N-2}) P(w_1, w_2, \ldots. w_{N-2})$$

$$\cdots$$

$$= P(w_N \mid w_1, w_2, \ldots w_{N-1}) P(w_{N-1} \mid w_1, w_2, \ldots w_{N-2}) \ldots P(w_2 \mid w_1) P(w_1)$$

$$= \prod_{i=1}^{N} P(w_i \mid w_1, w_2, w_3, \ldots\ldots w_{i-1})$$

Bayes Decomposition: write joint probability as product of conditional probabilities

In summary

$$P(w_1, w_2, w_3, \ldots\ldots w_N) =$$

$$= \prod_{i=1}^{N} P(w_i \mid w_1, w_2, w_3, \ldots\ldots w_{i-1})$$

Usage:
process sentences from left to right in speech recognition, machine translation, …!

Use Bayes Decomposition in Classification

For classification of documents in class c (e.g. a topic) you need to estimate

$$\mathrm{P}(\mathrm{w}_1, w_2, w_3, \ldots\ldots w_N \mid c)$$

Using the decomposition and a strong independence assumption this results in

$$\mathrm{P}(\mathrm{w}_1, w_2, w_3, \ldots\ldots w_N \mid c) = \prod_{i=1}^{N} \mathrm{P}(w_i \mid \mathrm{w}_1, w_2, w_3, \ldots\ldots w_{i-1}, c) \approx \prod_{i=1}^{N} \mathrm{P}(w_i \mid c)$$

This is the major building block of a Naïve Bayes classifier

Note: instead of words you can also use other features

Summary

- In the context of this lecture we have a wide definition of language
- Simple statistical analysis show similarity of
  - Natural languages, formal languages and descriptive languages
- Revision of probability theory