17 Stochastic Languages

The stochastic model of language as a stationary Markov process easily led to useful theoretic results that fit well with empirical observations. On the other hand it is far from the computer scientific model that regards a language as a fixed set of strings with certain properties and that is intuitively much closer to reality. In fact the Markov model may produce *every* string in Σ^* with a non-zero probability! (We assume that each letter $s \in \Sigma$ has a non-zero probability—otherwise we would throw it away.) Experience tells us that only a very small portion of all character strings represent meaningful texts in any natural language. Here we consider an alternative model that respects this facet of reality, but otherwise is somewhat cumbersome.

Recall from Chapter 1 that a language is a subset $M \subseteq \Sigma^*$.

A Computer Theoretic Model

The statistical cryptanalysis of the monoalphabetic substitution relied on the hypothesis—supported by empirical evidence—that the average relative frequencies of the letters $s \in \Sigma$ in texts of sufficient length from this language approximate typical values p_s. This is even true when we consider only fixed positions j in the texts, at least for almost all j—the first letters of texts for example usually have different frequencies.

Now we try to build a mathematical model of language that reflects this behaviour. Let $M \subseteq \Sigma^*$ a language, and $M_r := M \cap \Sigma^r$ for $r \in \mathbb{N}$ the set of texts of length r. The average frequency of the letter $s \in \Sigma$ at the position $j \in [0 \ldots r-1]$ of texts in M_r is

$$\mu_{sj}^{(r)} := \frac{1}{\#M_r} \cdot \sum_{a \in M_r} \delta_{sa_j}$$

(This sum counts the texts $a \in M_r$ with the letter s at position j.)

Example Let $M = \Sigma^*$ Then

$$\mu_{sj}^{(r)} := \frac{1}{n^r} \cdot \sum_{a \in \Sigma^r} \delta_{sa_j} = \frac{1}{n} \quad \text{for all } s \in \Sigma,\ j = 1, \ldots, r-1,$$

because there are exactly n^{r-1} possible texts with fixed $a_j = s$.

Definition

The language $M \subseteq \Sigma^*$ is called **stochastic** if there is at most a finite exceptional set $J \subseteq \mathbb{N}$ of positions such that

$$p_s := \lim_{r \to \infty} \mu_{sj}^{(r)}$$

exists uniformly in j and is independent from j for all $j \in \mathbb{N} - J$ and all $s \in \Sigma$.

The p_s are called the **letter frequencies** of M and obviously coincide with the limit values for the frequencies of the letters over the complete texts.

Examples and Remarks

1. The exceptional set J for natural languages usually consists only of the start position 0 and the end position. That is, the first and last letters of texts may have different frequencies. For example in English the letter "t" is the most frequent first letter instead of "e", followed by "a" and "o". In German this is "d", followed by "w", whereas "t" almost never occurs as first letter.
2. The language $M = \Sigma^*$ is stochastic.
3. Because always $\sum_{s\in\Sigma} \mu_{sj}^{(r)} = 1$, also $\sum_{s\in\Sigma} p_s = 1$.

Note that this notation is not standard in the literature.

Also note that we consider a *theoretical model.* For a natural language it may not be well-defined whether a given text is meaningful or not, not even if it is taken from a newspaper.

The Mean Coincidence Between Two Languages

Let $L, M \subseteq \Sigma^*$ two stochastic languages with letter frequencies q_s and p_s for $s \in \Sigma$. We consider the mean value of the coincidences of texts of length r:

$$\kappa_{LM}^{(r)} := \frac{1}{\#L_r} \cdot \frac{1}{\#M_r} \cdot \sum_{a\in L_r} \sum_{b\in M_r} \kappa(a,b)$$

Theorem 5 *The mean coincidence of the stochastic languages L and M is asymptotically*

$$\lim_{r\to\infty} \kappa_{LM}^{(r)} = \sum_{s\in\Sigma} p_s q_s$$

The proof follows.

Interpretation: The coincidence of sufficiently long texts of the same length is approximately

$$\kappa(a,b) \approx \sum_{s\in\Sigma} p_s q_s$$

An Auxiliary Result

Lemma 5 *Let M be a stochastic language. Then the average deviation for all letters $s \in \Sigma$*

$$\frac{1}{r} \cdot \sum_{j=0}^{r-1} \left(\mu_{sj}^{(r)} - p_s \right) \to 0 \quad \textit{for } r \to \infty$$

Proof. Fix $\varepsilon > 0$, and let r large enough that

1. $r \geq 4 \cdot \frac{\#J}{\varepsilon}$,
2. $|\mu_{sj}^{(r)} - p_s| < \frac{\varepsilon}{2}$ for all $j \in [0 \ldots r] - J$.

For $j \in J$ we have $|\mu_{sj}^{(r)} - p_s| \leq |\mu_{sj}^{(r)}| + |p_s| \leq 2$. Therefore

$$\frac{1}{r} \cdot \sum_{j=0}^{r-1} |\mu_{sj}^{(r)} - p_s| < \frac{1}{r} \cdot 2 \cdot \#J + \frac{r - \#J}{r} \cdot \frac{\varepsilon}{2} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

◇

Remark The mean frequency of s in texts of length r is

$$\mu_s^{(r)} = \frac{1}{r} \cdot \sum_{j=0}^{r-1} \mu_{sj}^{(r)} = \frac{1}{r} \cdot \frac{1}{\#M_r} \cdot \sum_{a \in M_r} \delta_{sa_j}$$

For this we get the limit

Corollary 5 $\lim_{r \to \infty} \mu_s^{(r)} = p_s$

Proof of the Theorem

$$\begin{aligned}
\kappa_{LM}^{(r)} &= \frac{1}{\#L_r \cdot \#M_r} \cdot \sum_{a \in L_r} \sum_{b \in M_r} \left(\frac{1}{r} \cdot \sum_{j=0}^{r-1} \sum_{s \in \Sigma} \delta_{sa_j} \delta_{sb_j} \right) \\
&= \sum_{s \in \Sigma} \frac{1}{r} \cdot \sum_{j=0}^{r-1} \left[\frac{1}{\#L_r} \sum_{a \in L_r} \delta_{sa_j} \right] \cdot \left[\frac{1}{\#M_r} \sum_{b \in M_r} \delta_{sb_j} \right] \\
&= \sum_{s \in \Sigma} \frac{1}{r} \cdot \sum_{j=0}^{r-1} [q_s + \varepsilon_{sj}] \cdot [p_s + \eta_{sj}] \\
&= \sum_{s \in \Sigma} \left[p_s q_s + \frac{p_s}{r} \cdot \sum_{j=0}^{r-1} \varepsilon_{sj} + \frac{q_s}{r} \cdot \sum_{j=0}^{r-1} \eta_{sj} + \frac{1}{r} \cdot \sum_{j=0}^{r-1} \varepsilon_{sj} \eta_{sj} \right]
\end{aligned}$$

The second and third summands converge to 0 by the lemma. The fourth converges to 0 because $|\varepsilon_{sj}\eta_{sj}| \leq 1$. Therefore the sum converges to $\sum_{s \in \Sigma} p_s q_s$. ◇