16 Modeling a Language by a Markov Process

For deriving theoretical results a common model of language is the interpretation of texts as results of Markov processes. This model was introduced by Shannon in his fundamental papers published after World War II.

If we look at letter frequencies only, we define a Markov process of order 0. If we also incorporate bigram frequencies into our model, it becomes a Markov process of order 1, if we include trigram frequencies, of order 2, and so on.

In this section we want to derive theoretical expectation values for κ, φ, and χ. For this the order of the Markov model is irrelevant.

Message Sources

Let the alphabet Σ be equipped with a probability distribution, assigning the probability p_s to the letter $s \in \Sigma$. In particular $\sum_{s\in\Sigma} p_s = 1$. We call (Σ, p) a **message source** and consider random variables X in Σ, that is mappings $X\colon \Omega \longrightarrow \Sigma$ where Ω is a finite probability space with probability measure P, such that $P(X^{-1}s) = p_s$ for all $s \in \Sigma$.

Picking a letter of Σ at random from the message source is modeled as evaluating $X(\omega)$ for some $\omega \in \Omega$. We calculate the expectation values of the Kronecker symbols for random variables $X, Y\colon \Omega \longrightarrow \Sigma$ and letters $s \in \Sigma$ where Y may belong to a message source (Σ, q) with a possibly different probability distribution $q = (q_s)_{s\in\Sigma}$:

$$\delta_{sX}(\omega) = \begin{cases} 1 & \text{if } X(\omega) = s \\ 0 & \text{otherwise} \end{cases} \qquad \delta_{XY}(\omega) = \begin{cases} 1 & \text{if } X(\omega) = Y(\omega) \\ 0 & \text{otherwise} \end{cases}$$

Lemma 4 (i) $\mathrm{E}(\delta_{sX}) = p_s$ *for all* $s \in \Sigma$.
(ii) *If* X *and* Y *are independent, then* $\mathrm{E}(\delta_{XY}) = \sum_{s\in\Sigma} p_s q_s$.
(ii) *If* X *and* Y *are independent, then* δ_{sX} *and* δ_{sY} *are independent.*

Proof. (i) Since δ takes only the values 0 and 1, we have

$$\mathrm{E}(\delta_{sX}) = 1 \cdot P(X^{-1}s) + 0 \cdot P(\Omega - X^{-1}s) = P(X^{-1}s) = p_s.$$

(ii) In the same way, using the independence of X and Y,

$$\begin{aligned} \mathrm{E}(\delta_{X,Y}) &= 1 \cdot P(\omega \mid X(\omega) = Y(\omega)) + 0 \cdot P(\omega \mid X(\omega) \neq Y(\omega)) \\ &= P(X = Y) = \sum_{s\in\Sigma} P(X^{-1}s \cap Y^{-1}s) \\ &= \sum_{s\in\Sigma} P(X^{-1}s) \cdot P(Y^{-1}s) = \sum_{s\in\Sigma} p_s q_s \end{aligned}$$

(iii) $\delta_{sX}^{-1}(1) = \{\omega | X(\omega) = s\} = X^{-1}s$, and $\delta_{sX}^{-1}(0) = \Omega - X^{-1}s$. The same for Y. The assertion follows because $P(X^{-1}s \cap Y^{-1}s) = P(X^{-1}s) \cdot P(Y^{-1}s)$. $\diamond$

Picking a random text of length r is modeled by evaluating an r-tuple of random variables at some ω. This leads to the following definition:

Definition. A **message** of length r from the message source (Σ, p) is a sequence $X = (X_1, \ldots, X_r)$ of random variables $X_1, \ldots, X_r \colon \Omega \longrightarrow \Sigma$ such that $P(X_i^{-1}s) = p_s$ for all $i = 1, \ldots, r$ and all $s \in \Sigma$.

Note. In particular the X_i are identically distributed. They are not necessarily independent.

The Coincidence Index of Message Sources

Definition. Let $Y = (Y_1, \ldots, Y_r)$ be another message of length r from a possibly different message source (Σ, q). Then the **coincidence index** of X and Y is the random variable

$$\mathsf{K}_{XY} \colon \Omega \longrightarrow \mathbb{R}$$

defined by

$$\mathsf{K}_{XY}(\omega) := \frac{1}{r} \cdot \#\{i = 1, \ldots, r \mid X_i(\omega) = Y_i(\omega)\} = \frac{1}{r} \cdot \sum_{i=1}^{r} \delta_{X_i(\omega), Y_i(\omega)}$$

We calculate its expectation under the assumption that each pair of X_i and Y_i is independent. From Lemma 4, using the additivity of E, we get

$$\mathrm{E}(\mathsf{K}_{XY}) = \frac{1}{r} \cdot \sum_{i=1}^{r} \mathrm{E}(\delta_{X_i, Y_i}) = \frac{1}{r} \cdot r \cdot \sum_{s \in \Sigma} p_s q_s = \sum_{s \in \Sigma} p_s q_s$$

independently of the length r. Therefore it is adequate to call this expectation the **coincidence index κ_{LM} of the two message sources** L, M. We have proven:

Theorem 2 *The coincidence index of two message sources $L = (\Sigma, p)$ and $M = (\Sigma, q)$ is*

$$\kappa_{LM} = \sum_{s \in \Sigma} p_s q_s$$

Now we are ready to calculate theoretical values for the "typical" coincidence indices of languages under the assumption that the model "message source" fits their real behaviour:

Example 1, random texts versus any language M: Here all $p_s = 1/n$, therefore $\kappa_{\Sigma *} = n \cdot \sum_{s \in \Sigma} 1/n \cdot q_s = 1/n$.

Example 2, English texts versus English: From Table 39 we get the value 0.0653.

Example 3, German texts versus German: The table gives 0.0758.

Example 4, English versus German: The table gives 0.0664.

Note that these theoretical values for the real languages differ slightly from the former empirical values. This is due to two facts:

- The model—as every mathematical model—is an approximation to the truth.
- The empirical values underly statistical variations and depend on the kind of texts that were evaluated.

The Cross-Product Sum of Message Sources

For a message $X = (X_1, \ldots, X_r)$ from a message source (Σ, p) we define the (relative) letter frequencies as random variables

$$\mathrm{M}_{sX} \colon \Omega \longrightarrow \mathbb{R}, \quad \mathrm{M}_{sX} = \frac{1}{r} \cdot \sum_{i=1}^{r} \delta_{sX_i},$$

or more explicitly,

$$\mathrm{M}_{sX}(\omega) = \frac{1}{r} \cdot \#\{i \mid X_i(\omega) = s\} \quad \text{for all } \omega \in \Omega.$$

We immediately get the expectation

$$\mathrm{E}(\mathrm{M}_{sX}) = \frac{1}{r} \cdot \sum_{i=1}^{r} \mathrm{E}(\delta_{sX_i}) = p_s.$$

Definition. Let $X = (X_1, \ldots, X_r)$ be a message from the source (Σ, p), and $Y = (Y_1, \ldots, Y_t)$, a message from the source (Σ, q). Then the **cross-product sum** of X and Y is the random variable

$$\mathsf{X}_{XY} \colon \Omega \longrightarrow \mathbb{R}, \quad \mathsf{X}_{XY} := \frac{1}{rt} \cdot \sum_{s \in \Sigma} \mathrm{M}_{sX} \mathrm{M}_{sY}.$$

Table 39: Calculating theoretical values for coincidence indices

Letter s	English p_s	German q_s	Square p_s^2	Square q_s^2	Product $p_s q_s$
A	0.082	0.064	0.006724	0.004096	0.005248
B	0.015	0.019	0.000225	0.000361	0.000285
C	0.028	0.027	0.000784	0.000729	0.000756
D	0.043	0.048	0.001849	0.002304	0.002064
E	0.126	0.175	0.015876	0.030625	0.022050
F	0.022	0.017	0.000484	0.000289	0.000374
G	0.020	0.031	0.000400	0.000961	0.000620
H	0.061	0.042	0.003721	0.001764	0.002562
I	0.070	0.077	0.004900	0.005929	0.005390
J	0.002	0.003	0.000004	0.000009	0.000006
K	0.008	0.015	0.000064	0.000225	0.000120
L	0.040	0.035	0.001600	0.001225	0.001400
M	0.024	0.026	0.000576	0.000676	0.000624
N	0.067	0.098	0.004489	0.009604	0.006566
O	0.075	0.030	0.005625	0.000900	0.002250
P	0.019	0.010	0.000361	0.000100	0.000190
Q	0.001	0.001	0.000001	0.000001	0.000001
R	0.060	0.075	0.003600	0.005625	0.004500
S	0.063	0.068	0.003969	0.004624	0.004284
T	0.091	0.060	0.008281	0.003600	0.005460
U	0.028	0.042	0.000784	0.001764	0.001176
V	0.010	0.009	0.000100	0.000081	0.000090
W	0.023	0.015	0.000529	0.000225	0.000345
X	0.001	0.001	0.000001	0.000001	0.000001
Y	0.020	0.001	0.000400	0.000001	0.000020
Z	0.001	0.011	0.000001	0.000121	0.000011
Sum	1.000	1.000	0.0653	0.0758	0.0664

To calculate its expectation we assume that each X_i is independent of all Y_j, and each Y_j is independent of all X_i. Under this assumption let us call the messages X and Y **independent**. Then from Lemma 4 and the formula

$$\mathsf{X}_{XY} := \frac{1}{rt} \cdot \sum_{s\in\Sigma} \sum_{i=1}^{r} \sum_{j=1}^{t} \delta_{sX_i}\delta_{sY_j}$$

we get

$$\mathrm{E}(\mathsf{X}_{XY}) = \frac{1}{rt} \cdot \sum_{s\in\Sigma} \sum_{i=1}^{r} \sum_{j=1}^{t} \mathrm{E}(\delta_{sX_i})\mathrm{E}(\delta_{sY_j}) = \sum_{s\in\Sigma} p_s q_s$$

again independently of the length r. Therefore we call this expectation the **cross-product sum χ_{LM} of the two message sources** L, M. We have proven:

Theorem 3 *The cross-product sum of two message sources $L = (\Sigma, p)$ and $M = (\Sigma, q)$ is*

$$\chi_{LM} = \sum_{s\in\Sigma} p_s q_s.$$

The Inner Coincidence Index of a Message Source

Let $X = (X_1, \ldots, X_r)$ be a message from a source (Σ, p). In analogy with Sections 10 and 14 we define the random variables

$$\Psi_X, \Phi_X \colon \Omega \longrightarrow \mathbb{R}$$

by the formulas

$$\Psi_X := \sum_{s\in\Sigma} M_{sX}^2, \qquad \Phi_X := \frac{r}{r-1} \cdot \Psi_x - \frac{1}{r-1}.$$

We try to calculate the expectation of Ψ_X first:

$$\begin{aligned} \Psi_X &= \frac{1}{r^2} \cdot \sum_{s\in\Sigma} \left(\sum_{i=1}^{r} \delta_{sX_i} \right)^2 \\ &= \frac{1}{r^2} \cdot \sum_{s\in\Sigma} \left(\sum_{i=1}^{r} \delta_{sX_i} + \sum_{i=1}^{r} \sum_{j\neq i} \delta_{sX_i}\delta_{sX_j} \right) \end{aligned}$$

since $\delta_{sX_i}^2 = \delta_{sX_i}$. Taking the expectation value we observe that for a sensible result we need the assumption that X_i and X_j are *independent* for $i \neq j$.

> In the language of Markov chains this means that we assume a Markov chain of order 0: The single letters of the messages from the source are independent from each other.

Under this assumption we get

$$\begin{aligned}
\mathrm{E}(\Psi_X) &= \frac{1}{r^2} \cdot \sum_{s\in\Sigma} \left(\sum_{i=1}^{r} p_s + \sum_{i=1}^{r} \sum_{j\neq i} \mathrm{E}(\delta_{sX_i})\mathrm{E}(\delta_{sX_j}) \right) \\
&= \frac{1}{r^2} \cdot \left(\sum_{i=1}^{r} \underbrace{\sum_{s\in\Sigma} p_s}_{1} + \sum_{s\in\Sigma} p_s^2 \cdot \underbrace{\sum_{i=1}^{r} \sum_{j\neq i} 1}_{r\cdot(r-1)} \right) \\
&= \frac{1}{r} + \frac{r-1}{r} \cdot \sum_{s\in\Sigma} p_s^2.
\end{aligned}$$

For Φ_X the formula becomes a bit more elegant:

$$\mathrm{E}(\Phi_X) = \frac{r}{r-1} \cdot \left(\frac{r-1}{r} \cdot \sum_{s\in\Sigma} p_s^2 + \frac{1}{r} \right) - \frac{1}{r-1} = \sum_{s\in\Sigma} p_s^2.$$

Let us call this expectation $\mathrm{E}(\Phi_X)$ the **(inner) coincidence index** of the message source (Σ, p), and let us call (by abuse of language) the message source **of order** 0 if its output messages are MARKOV chains of order 0 only. (Note that for a mathematically correct definition we should have included the "transition probabilities" into our definition of message source.) Then we have proved

Theorem 4 *The coincidence index of a message source $L = (\Sigma, p)$ of order 0 is*

$$\varphi_L = \sum_{s\in\Sigma} p_s^2.$$

The assumption of order 0 is relevant for small text lengths and negligeable for large texts, because for "natural" languages dependencies between letters affect small distances only. Reconsidering the tables in Section 11 we note in fact that the values for texts of lengths 100 correspond to the theoretical values, whereas for texts of lengths 26 the values are suspiciously smaller. An explanation could be that repeated letters, such as `ee`, `oo`, `rr`, are relatively rare and contribute poorly to the number of coincidences. This affects the power of the φ-test in an unfriendly way.

On the other hand considering SINKOV's test for the period in Section 13 we note that the columns of a polyalphabetic ciphertext are decimated excerpts from natural texts where the dependencies between letters are irrelevant: The assumption of order 0 is justified for SINKOV's test.