6 The Values of Bigram Scores

See the web page `http://www.staff.uni-mainz.de/pommeren/Cryptology/Classic/8_Transpos/cBLWsc.html`

Theoretical Values for Random Bigrams

Let $\Sigma = (s_0, \ldots, s_{n-1})$ be an alphabet and consider a probability distribution that assigns the probabilities p_i to the letters s_i. Choosing two letters independently from this distribution assigns the probability $p_i p_j$ to the bigram $s_i s_j$. Giving the bigrams whatever weights w_{ij} and scoring a set of bigrams by summing their weights the expected value of the weight of a bigram is

$$\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} w_{ij} p_i p_j.$$

Using this formula with the letter and bigram frequencies of natural languages and the corresponding conditional bigram log-weights we get the table

English: 1.47	German: 1.54	French: 1.48

Theoretical Values for True Bigrams

For a "true" bigram we first choose the first letter s_i with probability p_i, then we choose the second letter s_j with conditional probability $p_{j|i}$. This assigns the probability $p_i p_{j|i} = p_{ij}$ to the bigram $s_i s_j$, and the expected conditional bigram log-weight is

$$\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} w_{ij} p_{ij}.$$

Using this formula with the letter and bigram frequencies of natural languages and the corresponding conditional bigram log-weights we get the table

English: 1.94	German: 1.96	French: 1.99

Empirical Values for Natural Languages

See the web page `http://www.staff.uni-mainz.de/pommeren/Cryptology/Classic/8_Transpos/cBLWsc.html`