Maschinelles Übersetzen natürlicher Sprachen

11. Übungsblatt

2017-01-26

Aufgabe 1

We consider the CFG G with the set of productions R:

$$\rho_1 \colon S \to SS\,, \qquad\qquad \rho_2 \colon S \to a\,,$$

which generates the language $\{a^n \mid n \geq 1\}$. In this scenario, we let $X = \Sigma^* \cup \{\bot\}$, $Y = D(G) \cup \{\bot\}$ where $D(G)$ is the set of all (leftmost) derivations of G, and $Z = \{\emptyset\}$.

We fix the probability model $p \colon \Omega \to \mathcal{M}(Y \times X \mid Z)$ by letting $\Omega = [0,1]$, i.e., the interval of reals from 0 to 1. Intuitively, $\omega \in \Omega$ is the probability of production ρ_1, and hence $1-\omega$ is the probability of production ρ_2. Note that every derivation for a^n consists of $n-1$ occurrences of ρ_1 and n occurrences of ρ_2. Then, as usual, we define the probability distribution $p_\omega(\emptyset)$ of $Y \times X$ follows:

$$p_\omega(\emptyset)(y,x) = \begin{cases} \omega^{n-1} \cdot (1-\omega)^n & \text{if } y \text{ derives } x,\ x = a^n, \text{ and } n \geq 1, \\ 1 - \sum_{n \geq 1} C_{n-1} \cdot \omega^{n-1} \cdot (1-\omega)^n & \text{if } y = \bot \text{ and } x = \bot, \\ 0 & \text{otherwise,} \end{cases}$$

where C_n is the number of derivations for a^{n+1}, which is given by

$$C_n = \frac{(2n)!}{n! \cdot (n+1)!}\,. \qquad \text{(Catalan number)}$$

1. Let us consider the $X \times Z$-corpus c with

   $$c(aa,\emptyset) = 4, \;\; c(aaa,\emptyset) = 6, \quad \text{and } c(x,\emptyset) = 0 \text{ for every other } x.$$

   Derive the $Y \times X \times Z$-corpus $c\langle \omega, p\rangle$. Then compute $(\!|p|\!)_{\mathrm{cb}}(\omega)$.

2. We let $A = \{SS, a\}$ and $B = \{S\}$ be the sets of, respectively, all right-hand sides and all left-hand sides in the set R of productions. Moreover, we let $C = A \times B$. Define an appropriate counting information $\kappa = (q, \lambda, \pi)$ such that $p_\omega(y, x \mid \emptyset) = (\kappa^\flat)_\omega(y, x \mid \emptyset)$.

   Consider the corpus c of task 1 and specify the relevant entries of the corpus $c\langle \omega, \kappa\rangle$. Afterwards, give the simple counting step mapping $(\!|\kappa|\!)_{\mathrm{sc}}(\omega)$.

3. We define the io-info $\mu = (q, \pi_1, \pi_2, K, H)$ with q as before and

   - $\pi_1 \colon Y_{\not\bot} \to X_{\not\bot}$ maps every derivation to its derived string in Σ^*, and $\pi_2 \colon Y_{\not\bot} \to Z$ maps every derivation to $\emptyset$,
   - $K(\emptyset)$ is the unambiguous RTG with one state $*$ and the rules $\langle **, (SS, S), *\rangle$ and $\langle \varepsilon, (a, S), *\rangle$,
   - $H(a^n, \emptyset)$ is the unambiguous RTG with states $\{1, \ldots, n\}$, n being initial, and the rules $\langle jk, (SS, S), j+k\rangle$ and $\langle \varepsilon, (a, S), 1\rangle$.

Compute:

$$\chi_{\omega,aa,\emptyset}(SS,S) = \qquad \beta_{\omega,aa,\emptyset} =$$
$$\chi_{\omega,aa,\emptyset}(a,S) = \qquad \beta_{\omega,aaa,\emptyset} =$$
$$\chi_{\omega,aaa,\emptyset}(SS,S) = \qquad c\langle\omega,\mu\rangle(SS,S) =$$
$$\chi_{\omega,aaa,\emptyset}(a,S) = \qquad c\langle\omega,\mu\rangle(a,S) =$$

Show that $(\!|\mu|\!)_{\mathrm{io}}(\omega) \ni \widetilde{c\langle\omega,\mu\rangle}$ and compute the following values:

$$\widetilde{c\langle\omega,\mu\rangle}(SS,S) = \qquad \widetilde{c\langle\omega,\mu\rangle}(a,S) =$$