Slovní úloha

Zadání

Předpokládejme, že existují 2 jazyky - L_1 a L_2, které nemají žádné stejné slovo. Podmíněná entropie na textu T_1 pro jazyk L_1 je E a podmíněná entropie na textu T_2 pro jazyk L_2 je také E. Nový text T vznikne spojením textů T_1 a T_2. Bude podmíněná entropie tohoto nového textu větší, stejná nebo menší než ta původní?

Řešení

Základní vzorec pro výpočet podmíněné entropie vypadá následovně:

$$E = -\sum_{i,j \in T} p(i,j) \cdot \log(p(j|i))$$

Pro jazyk L_1 a text T_1 platí:

$$E = -\sum_{i_1, j_1 \in T_1} \frac{c(i_1, j_1)}{N_1} \cdot \log(\frac{c(i_1, j_1)}{c(i_1)})$$

kde $c(i_1, j_1)$ je počet výskytů dvojit slov i_1 a j_1, $c(i_1)$ je počet výskytů slova i_1 a N_1 je počet slov v textu T_1.

Pro jazyk L_2 a text T_2 platí:

$$E = -\sum_{i_2, j_2 \in T_2} \frac{c(i_2, j_2)}{N_2} \cdot \log(\frac{c(i_2, j_2)}{c(i_2)})$$

kde $c(i_2, j_2)$ je počet výskytů dvojit slov i_2 a j_2, $c(i_2)$ je počet výskytů slova i_2 a N_2 je počet slov v textu T_2.

Pro podmíněnou entropii E_m spojených textů T_1 a T_2 platí:

$$E_m = -\sum_{i_m, j_m \in T_1 \cup T_2} \frac{c(i_m, j_m)}{N_1 + N_2 + 1} \cdot \log(\frac{c(i_m, j_m)}{c(i_m)})$$

kde $c(i_m, j_m)$ je počet výskytů dvojit slov i_m a j_m, $c(i_m)$ je počet výskytů slova i_m a N_1 je počet slov v textu T_1 a N_2 je počet slov v textu T_2.

Tuto rovnici rozepíšeme pro jednotlivé texty T_1 a T_2 a hraniční bigram.

$$\begin{aligned} E_m = & -\sum_{i_1, j_1 \in T_1} \frac{c(i_1, j_1)}{N_1 + N_2 + 1} \cdot \log(\frac{c(i_1, j_1)}{c(i_1)}) \\ & -\sum_{i_2, j_2 \in T_2} \frac{c(i_2, j_2)}{N_1 + N_2 + 1} * \log(\frac{c(i_2, j_2)}{c(i_2)}) \\ & -\frac{c(i_1, j_2)}{N_1 + N_2 + 1} * \log(\frac{c(i_1, j_2)}{c(i_1)}) \end{aligned}$$

První suma téměř odpovídá vzorci pro podmíněnou entropii textu T_1, druhá suma téměř odpovídá vzorci pro podmíněnou entropii textu T_2. Poslední člen odpovídá výpočtu podmíněné entropie pro hraniční bigram. Člen $c(i_1, j_2)$ je roven 1 (texty mají disjunktní slovník) a $c(i_1)$ odpovídá počtu výskytů slova i_1 v textu T_1.
Po úpravě dostaneme následující výraz:

$$E_m = \frac{N_1 \cdot E}{N_1 + N_2 + 1} + \frac{N_2 \cdot E}{N_1 + N_2 + 1} + \frac{\log(c(i_1)) \cdot E}{N_1 + N_2 + 1}$$

Tento výraz srovnáme s původní hodnotou podmíněné entropie.

$$E \Leftrightarrow E_m$$

$$E \Leftrightarrow \frac{N_1 \cdot E}{N_1 + N_2 + 1} + \frac{N_2 \cdot E}{N_1 + N_2 + 1} + \frac{\log(c(i_1)) \cdot E}{N_1 + N_2 + 1}$$

$$\frac{N_1 \cdot E + N_1 \cdot E + E}{N_1 + N_2 + 1} \Leftrightarrow \frac{N_1 \cdot E + N_1 \cdot E + \log(c(i_1))}{N_1 + N_2 + 1}$$

$$N_1 \cdot E + N_1 \cdot E + E \Leftrightarrow N_1 \cdot E + N_1 \cdot E + \log(c(i_1))$$

Po úpravách dostaneme:

$$E \Leftrightarrow \log(c(i_1))$$

Pokud je $\log(c(i_1))$ menší než než původní podmíněná entropie E, tak podmíněná entropie sloučených textů bude menší.

Závěr

Vztah mezi hodnota podmíněná entropie samostatných textů a sloučeného textu je závislý na hodnotě podmíněné entropie E a počtu výskytů posledního slova z prvního textu T_1.

Pokud je $\log(c(i_1))$ menší než než původní podmíněná entropie E, tak podmíněná entropie sloučených textů bude menší.
Pokud je $\log(c(i_1))$ větší než než původní podmíněná entropie E, tak podmíněná entropie sloučených textů bude větší.
Jinak bude podmíněná entropie samostatných textů i sloučených textů shodná.