dvojice z odlišných textů

$$H = -\sum_{\substack{i\in I_1\\ j\in J_1}} p'(i,j)\cdot \log p'(j|i) - \sum_{\substack{i\in I_2\\ j\in J_2}} p'(i,j)\cdot \log p'(j|i) - \sum_{i\in I_1, j\in J_2} p'(i,j)\cdot \log p'(i|j)$$

$$- \sum_{i\in I_2, j\in J_1} p'(i,j)\cdot \log p'(i|j)$$

Sumu lze zjednodušit, a to tak, že jediné dvojice slov z odlišných textů se může vyskytovat na rozhraní. Označme tyto slova $x_i \in T_1$, $x_j \in T_2$

$$= -\sum_{\substack{i\in I_1\\ j\in J_1}} p'(i,j)\cdot \log p'(j|i) - \sum_{\substack{i\in I_2\\ j\in J_2}} p'(i,j)\cdot \log p'(j|i) - p(x_i,x_j)\cdot \log(p(x_j|x_i))$$

$$p'(i,j) = \frac{c(i,j)}{\#\text{dv. } T_1 + \#\text{dv. } T_2 + 1} \Rightarrow p'(i,j) \leq p(i,j)$$

$$p'(j|i) = p'(j,i)/p'(i) = \frac{c(j,i)}{\#\text{dv. } T_1 + \#\text{dv. } T_2 + 1} \cdot \frac{\#\text{slov. } T_1 + \#\text{slov } T_2}{c(i)}$$

$\#\text{slov} - 1 = \#\text{dvojic} \Rightarrow \#\text{slov} \sim \#\text{dvojic} \Rightarrow p'(j|i) \doteq \frac{c(j,i)}{c(i)}$

$\Leftrightarrow$ uvažujeme jen vyskytující se dvojice

Na rozmezí se mohou stát násl. situace ($x_i :=$ poslední slovo v T_1, $x_j :=$ první slovo v T_2)

- x_i se nachází pouze na konci T_1: $\forall i \in I_1: i \neq x_i$

  pak výpočet $-\sum_{\substack{i\in I_1\\ j\in J_1}} p'(i,j)\cdot \log p'(j|i)$ není nijak ovlivněn,

  zároveň $p(x_j|x_i) = 1 \Rightarrow \log(p(x_j|x_i)) = 0 \Rightarrow$ zelená část se nezmění, žlutá část bude 0.

- x_i se nachází i jinde v textu T_1: $\exists i \in I_1: i = x_i$

  pak ve výpočtu $-\sum_{\substack{i\in I_1\\ j\in J_1}} p'(i,j)\cdot \log p'(j|i)$ všechny hodnoty $p'(j|i)$ pro $i = x_i$,

  tedy $\log(p(j|i))$ se zvýší, potažmo i hodnota entropie.

  zároveň $p(x_j|x_i) < 1 \Rightarrow \log(p(x_j|x_i)) > 0 \Rightarrow$ zelená část se zvýší, žlutá část bude mít $\neq 0$ hodnotu

$$\Rightarrow = -\sum_{\substack{i\in I_1\\ j\in J_1}} p'(i,j)\cdot \log p(j|i) - \sum_{\substack{i\in I_2\\ j\in J_2}} p'(i,j)\cdot \log p(j|i) - p(x_j,x_i)\cdot \log(p(x_j|x_i))$$

$$p_1'(i,j) = \frac{\#\,dv.\,T_1}{\#dv.\,T_1 + \#dv.\,T_2}\cdot p_1(i,j) \qquad p_2'(i,j) = \frac{\#\,dv.\,T_2}{\#dv.\,T_1 + \#dv.\,T_2}\cdot p_2(i,j)$$

$$= -\sum_{\substack{i\in I_1\\ j\in J_1}} \frac{\#\,dv.\,T_1}{\#dv.\,T_1 + \#dv.\,T_2}\cdot p_1(i,j)\cdot \log p(j|i) - \sum_{\substack{i\in I_2\\ j\in J_2}} \frac{\#\,dv.\,T_2}{\#dv.\,T_1 + \#dv.\,T_2}\cdot p_2(i,j)\cdot \log p(j|i)$$

$$- p(x_j,x_i)\cdot \log(p(x_j|x_i))$$

$$= \frac{\#\,dv.\,T_1}{\#dv.\,T_1 + \#dv.\,T_2}\cdot E + \frac{\#\,dv.\,T_2}{\#dv.\,T_1 + \#dv.\,T_2}\cdot E - p(x_j,x_i)\cdot \log(p(x_j|x_i))$$

$$= E\cdot\left(\frac{\#\,dv.\,T_1}{\#dv.\,T_1 + \#dv.\,T_2} + \frac{\#\,dv.\,T_2}{\#dv.\,T_1 + \#dv.\,T_2}\right) - p(x_j,x_i)\cdot \log(p(x_j|x_i))$$

V případě, že žlutá část $=0$ (x_i se vyskytuje pouze na hraně T_1), je $H(T_1,T_2)=E$ a entropie se nezměnila.

V opačném případě v zelené části pro $i=x_i$:

$$p'(j|i) = \frac{c(j,i)}{c(i)+1} = \frac{c(j,i)}{c(i)}\cdot\frac{c(i)}{c(i)+1} = p(j|i)\cdot\frac{c(i)}{c(i)+1},$$

tedy zelená část se zvětší o $\frac{\#dv.\,v\,T_1}{\#dv.\,v\,T_1 + \#dv.\,v\,T_2}\cdot p(i,j)\cdot\log\left(p(j|i)\cdot\frac{1}{c(i)+1}\right) = A$

a do celkové sumy se přičítá i žlutý člen

$$-p(x_j,x_i)\cdot\log(p(x_j|x_i)) = B$$

Proto pokud: $(A-B) > 0 \Rightarrow$ entropie se zvýší

$(A-B) < 0 \Rightarrow$ entropie se sníží

Co bych měl ještě udělat:

Uniform prob. by měla vracet větší slovník než jen ten trénovací (protože v reálu bude více neviděných slov)

Takže něco jako uniform prob := $\frac{1}{|T| + \alpha(|H - T \cap H|)}$ $\alpha \geq 1$

Tohle α je potřeba si vyzkoušet z pusta.

T... train vocab.
H... herbert vocab

$H - T \cap H$ → počet „nových" slov, která nebyly v training datech.