Вывод формул ЕМ-алогоритма для задачи выравнивания переводов предложений

Дмитриев Леонид

20 мая 2023 г.

Содержание

Определения **2**

Первая модель **2**

Правдоподобие предложения . . . 2
Нижняя оценка логарифма правдоподобия . . . 2
Итоговая нижняя оценка . . . 3
Е-шаг . . . 3
Итоговое апостериорное распределение латентных переменных . . . 3
М-шаг . . . 3
Постановка оптимизационной задачи . . . 3
Функция Лагранжа . . . 3
Необходимые условия оптимальной пары $\Theta^\star$ и $\lambda^\star$ . . . 4
Решение системы . . . 4

Вторая модель **4**

Правдоподобие предложения . . . 4
Нижняя оценка логарифма правдоподобия . . . 4
Итоговая нижняя оценка . . . 5
Е-шаг . . . 5
Итоговое апостериорное распределение латентных переменных . . . 5
М-шаг . . . 5
Постановка оптимизационной задачи . . . 5
Функция Лагранжа . . . 6
Необходимые условия оптимальности $\Theta^\star$, $\Phi^\star$ и $\lambda^\star$ . . . 6
Решение системы . . . 6

Определения

- **A** - список из R векторов латентных переменных (истинных переводов слов целевого языка).
- **T** - список из R векторов слов целевого языка.
- **S** - список из R векторов слов исходного языка.
- m_k - длина вектора T_k.
- n_k - длина вектора S_k.
- Θ - матрица параметров модели $\in \mathbb{R}^{h\times l}$
- **h** - размер словаря исходного языка.
- **l** - размер словаря целевого языка.
- $\Theta_{xy} = P(y|x)$ - вероятность того, что переводом слова x с исходного языка на целевой является слово y.
- $q(A)$ - распределение латентных переменных.
- $q_k(A_k)$ - распределение латентных переменных для k-ой пары предложений.
- $q_{ki}(A_{ki})$ - распределение латентных переменных для i слова из целевого предложения k-ой пары предложений.
- $\Phi_{mn}(j|i) = p(a_i = j|m, n)$ - вероятность того, что в паре предложений с длинами m, n j-ому слову из целевого предложения будет выровнено i-ое слово из исходного.

Первая модель

Правдоподобие

Правдоподобие латентных переменных и предложения на целевом языке в этой модели записывается так:

$$p(A_k, T_k|S_k, \Theta) = \prod_{i=1}^{m_k} p(A_{ki})p(T_{ki}|A_{ki}, S_k, \Theta) = \prod_{i=1}^{m_k} \frac{1}{n_k}\theta(T_{ki}|S_{kA_{ki}}).$$

Нижняя оценка логарифма правдоподобия

$$\mathbb{E}_{q(A)} \log \frac{P(A,T|S,\Theta)}{q(A)} = \mathbb{E}_{q(A)} \log P(A,T|S,\Theta) - \mathbb{E}_{q(A)} \log q(A)$$

$\mathbb{E}_{q(A)} \log P(A,T|S,\Theta) = \mathbb{E}_{q(A)} \log \prod_{k=1}^{R} P(A_k,T_k|S_k,\Theta) = \mathbb{E}_{q(A)} \sum_{k=1}^{R} \log P(A_k,T_k|S_k,\Theta) = \sum_{k=1}^{R} \mathbb{E}_{q_k(A_k)} \log P(A_k,T_k|S_k,\Theta) = \sum_{k=1}^{R} \mathbb{E}_{q_k(A_k)} \log \prod_{i=1}^{m_k} P(A_{ki},T_{ki}|S_k,\Theta) = \sum_{k=1}^{R} \mathbb{E}_{q_k(A_k)} \sum_{i=1}^{m_k} \log P(A_{ki},T_{ki}|S_k,\Theta) = \sum_{k=1}^{R}\sum_{i=1}^{m_k} \mathbb{E}_{q_{ki}(A_{ki})} \log P(A_{ki},T_{ki}|S_k,\Theta) = \sum_{k=1}^{R}\sum_{i=1}^{m_k} \mathbb{E}_{q_{ki}(A_{ki})} \log P(A_{ki})P(T_{ki}|A_{ki},S_k,\Theta) = \sum_{k=1}^{R}\sum_{i=1}^{m_k} \mathbb{E}_{q_{ki}(A_{ki})} \log \frac{1}{n_k} P(T_{ki}|A_{ki},S_k,\Theta) = \sum_{k=1}^{R}\sum_{i=1}^{m_k} \mathbb{E}_{q_{ki}(A_{ki})} (\log \frac{1}{n_k} + \log P(T_{ki}|A_{ki},S_k,\Theta)) = -\sum_{k=1}^{R} m_k \log n_k + \sum_{k=1}^{R}\sum_{i=1}^{m_k} \mathbb{E}_{q_{ki}(A_{ki})} \log P(T_{ki}|A_{ki},S_k,\Theta)$

$\sum_{k=1}^{R}\sum_{i=1}^{m_k} \mathbb{E}_{q_{ki}(A_{ki})} \log P(T_{ki}|S_{kA_{ki}},\Theta) = \sum_{k=1}^{R}\sum_{i=1}^{m_k} \mathbb{E}_{q_{ki}(A_{ki})} \log P(T_{ki}|S_{kA_{ki}},\Theta) = \sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t) \log P(T_{ki}|S_{kt},\Theta) = \sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t) \log \Theta(T_{ki}|S_{kt})$

$$\mathbb{E}_{q(A)} \log q(A) = \sum_{k=1}^{R}\sum_{i=1}^{m_k} \mathbb{E}_{q_{ki}(A_{ki})} \log q_{ki}(A_{ki}) = \sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t) \log q_{ki}(t)$$

Итоговая нижняя оценка

$$\mathbb{E}_{q(A)} \log \frac{P(A, T|S, \Theta)}{q(A)} = \sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t) \log \Theta(T_{ki}|S_{kt}) - \sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t) \log q_{ki}(t) - \sum_{k=1}^{R} m_k \log n_k$$

Е-шаг

$$q^{\star}_{ki}(t) = P(t|T, S, \Theta) = P(t|T_{ki}, S_k, \Theta)$$

$$P(t|T_{ki}, S_k, \Theta) = \frac{P(t,T_{ki}|S_k,\Theta)}{P(T_{ki}|S_k,\Theta)} = \frac{P(t,T_{ki}|S_k,\Theta)}{\sum_{z=1}^{n_k} P(z,T_{ki}|S_k,\Theta)} = \frac{P(t)P(T_{ki}|t,S_k,\Theta)}{\sum_{z=1}^{n_k} P(z)P(T_{ki}|z,S_k,\Theta)} = \frac{P(T_{ki}|t,S_k,\Theta)}{\sum_{z=1}^{n_k} P(T_{ki}|z,S_k,\Theta)} = \frac{P(T_{ki}|S_{kt},\Theta)}{\sum_{z=1}^{n_k} P(T_{ki}|S_{kz},\Theta)} = \frac{\Theta(T_{ki}|S_{kt})}{\sum_{z=1}^{n_k} \Theta(T_{ki}|S_{kz})}$$

Итоговое апостериорное распределение латентных переменных

$$q^{\star}_{ki}(t) = \frac{\Theta(T_{ki}|S_{kt})}{\sum_{z=1}^{n_k} \Theta(T_{ki}|S_{kz})}$$

М-шаг

Постановка оптимизационной задачи

Оптимизируем по параметрам Θ, поэтому отбрасываем независящие от Θ слагаемые. Так как будет использован метод множителей Лагранжа для задачи условной минимизации, а наша задача максимизировать нижнюю оценку логарифма правдоподобия, изменим знак функционала, домножив его на -1.

$$\mathbb{J}(\Theta) = -\sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t) \log \Theta(T_{ki}|S_{kt}) \to min$$

$$g_{xy}(\Theta) = -\Theta_{xy} \leq 0, \quad (x, y) \in [1, h] \times [1, l]$$

$$g_x(\Theta) = \sum_{z=1}^{l} \Theta_{xz} - 1 = 0, \quad x \in [1, h]$$

Функция Лагранжа

Условие Слейтера выполняется (например для случая равномерного распределения вдоль оси целевого языка). Значит $\lambda_0 \neq 0$, поэтому можем нормализовать лямбды так, чтобы $\lambda_0 = 1$.

$$\mathbb{L}(\Theta, \lambda) = -\sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t) \log \Theta(T_{ki}|S_{kt}) - \sum_{x=1}^{h}\sum_{y=1}^{l} \lambda_{xy}\Theta_{xy} + \sum_{x=1}^{h} \lambda_x (\sum_{z=1}^{l} \Theta_{xz} - 1)$$

Необходимые условия оптимальной пары $\Theta^\star$ и $\lambda^\star$

$$\frac{\partial \mathbb{L}(\Theta^\star, \lambda^\star)}{\partial \Theta} = 0$$

$$\lambda^\star_{xy} >= 0, \quad (x, y) \in [1, h] \times [1, l], \quad \lambda^\star \neq \theta$$

$$\lambda^\star_{xy} * g_{xy}(\Theta^\star) = 0, \quad (x, y) \in [1, h] \times [1, l], \quad \lambda^\star_x \star g_x(\Theta^\star) = 0, \quad x \in [1, h]$$

Решение системы

$\frac{\partial \mathbb{L}(\Theta,\lambda)}{\partial \Theta_{xy}} = -\sum_{k=1}^{R} \sum_{i=1}^{m_k} \sum_{t=1}^{n_k} [S_{kt} == x][T_{ki} == y] q_{ki}(t) \frac{1}{\Theta(T_{ki}|S_{kt})} - \lambda_{xy} + \lambda_x =$
$= -\frac{1}{\Theta_{xy}} \sum_{k=1}^{R} \sum_{i=1}^{m_k} \sum_{t=1}^{n_k} [S_{kt} == x][T_{ki} == y] q_{ki}(t) - \lambda_{xy} + \lambda_x$

$$\text{Пусть} \quad K_{xy} = \sum_{k=1}^{R} \sum_{i=1}^{m_k} \sum_{t=1}^{n_k} [S_{kt} == x][T_{ki} == y] q_{ki}(t),$$

$$\text{тогда} \frac{\partial \mathbb{L}(\Theta, \lambda)}{\partial \Theta_{xy}} = \frac{K_{xy}}{\Theta_{xy}} - \lambda_{xy} + \lambda_x$$

Для $(x, y) \in [1, h] \times [1, l] \Rightarrow$

$$\lambda^\star_{xy} * \Theta^\star_{xy} = 0, \quad \frac{\partial \mathbb{L}(\Theta^\star, \lambda^\star)}{\partial \Theta_{xy}} = -\frac{K_{xy}}{\Theta^\star_{xy}} - \lambda^\star_{xy} + \lambda^\star_x = 0$$

$$\Rightarrow \lambda^\star_{xy} = 0, \quad \Theta^\star_{xy} = \frac{K_{xy}}{\lambda^\star_x}$$

Для $x \in [1, h] \Rightarrow$

$$\sum_{z=1}^{l} \Theta^\star_{xz} = \sum_{z=1}^{l} \frac{K_{xz}}{\lambda^\star_x} = \frac{1}{\lambda^\star_x} \sum_{z=1}^{l} K_{xz} = 1 \Rightarrow \lambda^\star_x = \sum_{z=1}^{l} K_{xz}$$

В итоге:

$$\Theta^\star_{xy} = \frac{K_{xy}}{\sum_{z=1}^{l} K_{xz}}, \quad (x, y) \in [1, h] \times [1, l],$$

$$\text{где} \quad K_{xy} = \sum_{k=1}^{R} \sum_{i=1}^{m_k} \sum_{t=1}^{n_k} [S_{kt} == x][T_{ki} == y] q_{ki}(t)$$

Вторая модель

Правдоподобие предложения

$$p(A_k, T_k|S_k, \Theta) = \prod_{i=1}^{m_k} p(A_{ki}|m, n) p(T_{ki}|A_{ki}, S_k, \Theta) = \prod_{i=1}^{m_k} \phi_{mn}(A_{ki}|i) \theta(T_{ki}|S_{kA_{ki}}).$$

Нижняя оценка логарифма правдоподобия

$$\mathbb{E}_{q(A)} \log \frac{P(A, T|S, \Theta)}{q(A)} = \mathbb{E}_{q(A)} \log P(A, T|S, \Theta) - \mathbb{E}_{q(A)} \log q(A)$$

$\mathbb{E}_{q(A)} \log P(A, T|S, \Theta) = \mathbb{E}_{q(A)} \log \prod_{k=1}^{R} P(A_k, T_k|S_k, \Theta) = \mathbb{E}_{q(A)} \sum_{k=1}^{R} \log P(A_k, T_k|S_k, \Theta) =$
$\sum_{k=1}^{R} \mathbb{E}_{q_k(A_k)} \log P(A_k, T_k|S_k, \Theta) = \sum_{k=1}^{R} \mathbb{E}_{q_k(A_k)} \log \prod_{i=1}^{m_k} P(A_{ki}, T_{ki}|S_k, \Theta) =$
$\sum_{k=1}^{R} \mathbb{E}_{q_k(A_k)} \sum_{i=1}^{m_k} \log P(A_{ki}, T_{ki}|S_k, \Theta) = \sum_{k=1}^{R} \sum_{i=1}^{m_k} \mathbb{E}_{q_{ki}(A_{ki})} \log P(A_{ki}, T_{ki}|S_k, \Theta) =$

$\sum_{k=1}^{R}\sum_{i=1}^{m_k}\mathbb{E}_{q_{ki}(A_{ki})}\log P(A_{ki}|m,n)P(T_{ki}|A_{ki},S_k,\Theta) =$
$\sum_{k=1}^{R}\sum_{i=1}^{m_k}\mathbb{E}_{q_{ki}(A_{ki})}\log P(A_{ki}|m,n) + \sum_{k=1}^{R}\sum_{i=1}^{m_k}\mathbb{E}_{q_{ki}(A_{ki})}\log P(T_{ki}|A_{ki},S_k,\Theta) =$
$\sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t)\log\Phi_{m_k n_k}(t|i) + \sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t)\log\Theta(T_{ki}|S_{kt})$

$$\mathbb{E}_{q(A)}\log q(A) = \sum_{k=1}^{R}\sum_{i=1}^{m_k}\mathbb{E}_{q_{ki}(A_{ki})}\log q_{ki}(A_{ki}) = \sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t)\log q_{ki}(t)$$

Итоговая нижняя оценка

$$\mathbb{E}_{q(A)}\log\frac{P(A,T|S,\Theta)}{q(A)} = \sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t)\log\Theta(T_{ki}|S_{kt}) + \sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t)\log\Phi_{m_k n_k}(t|i) - \sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t)\log q_{ki}(t)$$

$$\mathbb{E}_{q(A)}\log\frac{P(A,T|S,\Theta)}{q(A)} = \sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t)\log\frac{\Phi_{m_k n_k}(t|i)\Theta(T_{ki}|S_{kt})}{q_{ki}(t)}$$

E-шаг

$$q^{\star}_{ki}(t) = P(t|T,S,\Theta) = P(t|T_{ki},S_k,\Theta)$$

$P(t|T_{ki},S_k,\Theta) = \frac{P(t,T_{ki}|S_k,\Theta)}{P(T_{ki}|S_k,\Theta)} = \frac{P(t,T_{ki}|S_k,\Theta)}{\sum_{z=1}^{n_k}P(z,T_{ki}|S_k,\Theta)} = \frac{P(t|m_k,n_k)P(T_{ki}|t,S_k,\Theta)}{\sum_{z=1}^{n_k}P(z|m_k,n_k)P(T_{ki}|z,S_k,\Theta)} =$
$\frac{\Phi_{m_k n_k}(t|i)\Theta(T_{ki}|S_{kt})}{\sum_{z=1}^{n_k}\Phi_{m_k n_k}(z|i)\Theta(T_{ki}|S_{kz})}$

Итоговое апостериорное распределение латентных переменных

$$q^{\star}_{ki}(t) = \frac{\Phi_{m_k n_k}(t|i)\Theta(T_{ki}|S_{kt})}{\sum_{z=1}^{n_k}\Phi_{m_k n_k}(z|i)\Theta(T_{ki}|S_{kz})}$$

M-шаг

Постановка оптимизационной задачи

Оптимизируем по параметрам Θ и Φ, поэтому отбрасываем независящие от них слагаемые. Так как будет использован метод множителей Лагранжа для задачи условной минимизации, а наша задача максимизировать нижнюю оценку логарифма правдоподобия, изменим знак функционала, домножив его на -1.

$$\mathbb{J}(\Theta,\Phi) = -\sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t)\log\Theta(T_{ki}|S_{kt}) - \sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t)\log\Phi_{m_k n_k}(t|i) \to min$$

$$g_{xy}(\Theta,\Phi) = -\Theta_{xy} \le 0, \quad (x,y)\in[1,h]\times[1,l]$$

$$g_{x}(\Theta,\Phi) = \sum_{z=1}^{l}\Theta_{xz} - 1 = 0, \quad x\in[1,h]$$

$$g^{mn}_{ij}(\Theta,\Phi) = -\Phi^{mn}_{ij} \le 0, \quad (m,n) - \text{возможные пары длин предложений в корпусе}, \quad (i,j)\in[1,m]\times[1,n]$$

$$g_i^{mn}(\Theta,\Phi)=\sum_{z=1}^{n}\Phi_{iz}^{mn}-1=0,\quad (m,n)-\text{возможные пары длин предложений в корпусе},\quad i\in[1,m]$$

Функция Лагранжа

Условие Слейтера выполняется (например для случая равномерного распределения вдоль оси целевого языка в матрице Θ и равномерного распределения во всех матрицах Φ^{mn} вдоль оси исходного языка). Значит $\lambda_0 \neq 0$, поэтому можем нормализовать лямбды так, чтобы $\lambda_0 = 1$.

$\mathbb{L}(\Theta,\Phi,\lambda) = -\sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t)\log\Theta(T_{ki}|S_{kt}) - \sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k} q_{ki}(t)\log\Phi_{m_k n_k}(t|i) - \sum_{x=1}^{h}\sum_{y=1}^{l}\lambda_{xy}\Theta_{xy} + \sum_{x=1}^{h}\lambda_x(\sum_{z=1}^{l}\Theta_{xz}-1) - \sum_{(m,n)}\sum_{i=1}^{m}\sum_{j=1}^{n}\lambda_{ij}^{mn}\Phi_{ij}^{mn} + \sum_{(m,n)}\lambda_i^{mn}(\sum_{z=1}^{n}\Phi_{iz}^{mn}-1)$

Необходимые условия оптимальности $\Theta^\star$, $\Phi^\star$ и $\lambda^\star$

$$\frac{\partial\mathbb{L}(\Theta^\star,\Phi^\star,\lambda^\star)}{\partial\Theta}=\theta$$

$$\frac{\partial\mathbb{L}(\Theta^\star,\Phi^\star,\lambda^\star)}{\partial\Phi}=\theta$$

$$\lambda^\star\neq\theta$$

$$\lambda_{xy}^\star\geq 0,\quad (x,y)\in[1,h]\times[1,l]$$

$$\lambda_{xy}^\star * g_{xy}(\Theta^\star,\Phi^\star)=0,\quad (x,y)\in[1,h]\times[1,l]$$

$$\lambda_x^\star \star g_x(\Theta^\star,\Phi^\star)=0,\quad x\in[1,h]$$

$$\lambda^{\star mn}_{ij}\geq 0,\quad (m,n)-\text{возможные пары длин предложений в корпусе},\quad (i,j)\in[1,m]\times[1,n]$$

$$\lambda^{\star mn}_{ij} * g_{ij}^{mn}(\Theta^\star,\Phi^\star)=0,\quad (m,n)-\text{возможные пары длин предложений в корпусе},\quad (i,j)\in[1,m]\times[1,n]$$

$$\lambda^{\star mn}_{i} \star g_{i}^{mn}(\Theta^\star,\Phi^\star)=0,\quad (m,n)-\text{возможные пары длин предложений в корпусе},\quad i\in[1,m]$$

Решение системы

$\frac{\partial\mathbb{L}(\Theta,\lambda)}{\partial\Theta_{xy}} = -\sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k}[S_{kt}==x][T_{ki}==y]q_{ki}(t)\frac{1}{\Theta(T_{ki}|S_{kt})} - \lambda_{xy} + \lambda_x =$
$= -\frac{1}{\Theta_{xy}}\sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k}[S_{kt}==x][T_{ki}==y]q_{ki}(t) - \lambda_{xy} + \lambda_x = \frac{K_{xy}}{\Theta_{xy}} - \lambda_{xy} + \lambda_x$

$$K_{xy}=\sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k}[S_{kt}==x][T_{ki}==y]q_{ki}(t)$$

$\frac{\partial\mathbb{L}(\Theta,\lambda)}{\partial\Phi_{ij}^{mn}} = -\frac{1}{\Phi_{ij}^{mn}}\sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k}[m_k==m][n_k==n]q_{ki}(t) - \lambda_{ij}^{mn} + \lambda_i^{mn} = -\frac{F_{mn}}{\Phi_{ij}^{mn}} - \lambda_{ij}^{mn} + \lambda_i^{mn}$

$$F_{mn}=\sum_{k=1}^{R}\sum_{i=1}^{m_k}\sum_{t=1}^{n_k}[m_k==m][n_k==n]q_{ki}(t)$$

Для $(x,y)\in[1,h]\times[1,l]\Rightarrow$

$$\lambda_{xy}^\star * \Theta_{xy}^\star = 0,\quad \frac{\partial\mathbb{L}(\Theta^\star,\lambda^\star)}{\partial\Theta_{xy}} = -\frac{K_{xy}}{\Theta_{xy}^\star} - \lambda_{xy}^\star + \lambda_x^\star = 0$$

$$\Rightarrow \lambda^\star_{xy} = 0, \quad \Theta^\star_{xy} = \frac{K_{xy}}{\lambda^\star_x}$$

Для $x \in [1, h] \Rightarrow$

$$\sum_{z=1}^{l} \Theta^\star_{xz} = \sum_{z=1}^{l} \frac{K_{xz}}{\lambda^\star_x} = \frac{1}{\lambda^\star_x} \sum_{z=1}^{l} K_{xz} = 1 \Rightarrow \lambda^\star_x = \sum_{z=1}^{l} K_{xz}$$

Аналогично для (m, n) – возможные пары длин предложений в корпусе, $\quad (i, j) \in [1, m] \times [1, n] \Rightarrow$

$$\lambda^{\star}{}^{mn}_{ij} = 0, \quad \lambda^{\star}{}^{mn}_{i} = \sum_{z=1}^{n} F_{mz}, \quad \Phi^{mn}_{ij} = \frac{F_{mn}}{\lambda^{\star}{}^{mn}_{i}}$$

В итоге:

$$\Theta^\star_{xy} = \frac{K_{xy}}{\sum_{z=1}^{l} K_{xz}}, \quad (x, y) \in [1, h] \times [1, l],$$

$$\text{где} \quad K_{xy} = \sum_{k=1}^{R} \sum_{i=1}^{m_k} \sum_{t=1}^{n_k} [S_{kt} == x][T_{ki} == y] q_{ki}(t)$$

Для (m, n) – возможные пары длин предложений в корпусе, $\quad (i, j) \in [1, m] \times [1, n] \Rightarrow$

$$\Phi^{mn}_{ij} = \frac{F_{mn}}{\sum_{z=1}^{n} F_{mz}},$$

$$\text{где} \quad F_{mn} = \sum_{k=1}^{R} \sum_{i=1}^{m_k} \sum_{t=1}^{n_k} [m_k == m][n_k == n] q_{ki}(t)$$