EM-алгоритм для задачи машинного перевода

Пусть $S = (s_1, \dots, s_n)$ исходное предложение, $T = (t_1, \dots, t_m)$ — его перевод. В роли латентных переменных будут выступать выравнивания $A = (a_1, \dots, a_m)$ каждого слова в целевом предложении, причём $a_i \in \{1, \dots, n\}$ (считаем, что каждое слово в t является переводом какого-то слова из s). Параметрами модели является матрица условных вероятностей перевода: каждый её элемент $\theta(y|x) = p(y|x)$ отражает вероятность того, что переводом слова x с исходного языка на целевой является слово y (нормировка, соответственно, совершается по словарю целевого языка). Правдоподобие латентных переменных и предложения на целевом языке в этой модели записывается так:

$$p(A, T|S) = \prod_{i=1}^{m} p(a_i)p(t_i|a_i, S) = \prod_{i=1}^{m} \frac{1}{n}\theta(t_i|s_{a_i}).$$

Обозначим за V^s - словарь исходного языка, а за V^t словарь целевого.

EM-алгоритм

В EM-алгорите мы макисимизируем $p(T|S)$. Причём по всем парам предложений из нашего корпуса данных. Пусть таких пар всего R штук и длины в каждой r-ой паре равны n_r и m_r для исходного и целевого языка соответственно. Так как log - монотонная функция, то исходная задача эквивалентна максимизации лог-правдободобия. Для начала получим выражение для подсчёта нижней оценки правдоподобия. Для некторого распределения $q(A)$ на скрытых переменных верно:

$$log(\prod_{r=1}^{R} p(T^r|S^r)) = \sum_{r=1}^{R} \int q(A^r) log\, p(T^r|S^r) dA^r =$$

$$= \sum_{r=1}^{R} \int q(A^r) log \frac{p(A^r, T^r|S^r)}{p(A^r|T^r, S^r)} dA^r = \sum_{r=1}^{R} \int q(A^r) log \frac{p(A^r, T^r|S^r) q(A^r)}{p(A^r|T^r, S^r) q(A^r)} dA^r =$$

$$= \sum_{r=1}^{R} \int q(A^r) log \frac{p(A^r, T^r|S^r)}{q(A^r)} dA^r + \sum_{r=1}^{R} \int q(A^r) log \frac{q(A^r)}{p(A^r|T^r, S^r)} dA^r =$$

$$= \mathcal{L}(q, \theta) + KL(q||p)$$

Будем считать, что $q(A^r) = p(A^r|T^r, S^r)$. Тогда $KL(q||p) = 0$. Также учтём, что в данной задаче $p(A^r|T^r, S^r)$ имеет дискретное распределение, поэтому интегралы перейдут в суммы. Таким образом получим, что:

$$\mathcal{L} = \sum_{r=1}^{R} \sum_{i=1}^{m_r} \sum_{j=1}^{n_r} p(a_i = j \mid T^r, S^r) \; log \; \frac{\theta(t_i^r | s_{a_i^r}^r)}{n_r p(a_i = j \mid T^r, S^r)}$$

E-шаг

Мы хотим получить апостериорное распределение латентных переменных $p(A, |T, S)$ (для каждого предложения в нашем корпусе данных). Тогда по формуле Байеса получим:

$$p(a_i = j \mid T, S) = \frac{p(T|a_i = j, S) p(a_i = j|S)}{p(T|S)} = \frac{p(T, a_i = j|S)}{p(T|S)}$$

Таким образом, мы получили, что апостериорная вероятность представляет собой сумму вероятностей всех выравниваний, содержащих связь между t_i и s_j ($a_i = j$), деленную на сумму вероятностей всех возможных выравниваний:

$$\frac{p(T, a_i = j|S)}{p(T|S)} = \frac{\sum\limits_{A:a_i=j} p(A,T|S)}{\sum\limits_{A} p(A,T|S)} = \frac{\sum\limits_{A:a_i=j} \prod\limits_{k=1}^{m} \frac{1}{n}\theta(t_k|s_{a_k})}{\sum\limits_{A} \prod\limits_{k=1}^{m} \frac{1}{n}\theta(t_k|s_{a_k})} =$$

$$= \frac{\theta(t_i|s_j) \sum\limits_{a_1=1}^{n} ... \sum\limits_{a_{i-1}=1}^{n} \sum\limits_{a_{i+1}=1}^{n} ... \sum\limits_{a_m=1}^{n} \prod\limits_{k=1,k\neq i}^{m} \theta(t_k|s_{a_k})}{\sum\limits_{a_1=1}^{n} ... \sum\limits_{a_m=1}^{n} \prod\limits_{k=1}^{m} \theta(t_k|s_{a_k})} =^*$$

$$=^* \frac{\theta(t_i|s_j) \prod\limits_{k=1,k\neq i}^{m} \sum\limits_{a_1=1}^{n} ... \sum\limits_{a_{i-1}=1}^{n} \sum\limits_{a_{i+1}=1}^{n} ... \sum\limits_{a_m=1}^{n} \theta(t_k|s_{a_k})}{\prod\limits_{k=1}^{m} \sum\limits_{a_1=1}^{n} ... \sum\limits_{a_m=1}^{n} \theta(t_k|s_{a_k})} = \frac{\theta(t_i|s_j)}{\sum\limits_{a_i=1}^{n} \theta(t_i|s_{a_i})}$$

Таким образом, получим апостериорные вероятности для скрытых переменных для каждой пары предложений:

$$p(a_i = j \mid T, S) = \frac{\theta(t_i|s_j)}{\sum\limits_{k=1}^{n} \theta(t_i|s_k)}$$

M-шаг

На M-шаге мы максимизируем правдободобие $p(A,T|S)$. Причём по всем парам предложений из нашего корпуса данных. Тогда получим следующую оптимизационную задачу:

$$\mathbb{E}_A \, log(\prod_{r=1}^{R} p(A^r, T^r|S^r)) \to \max_{\theta}$$

Выведем формулы для обновления наших параметров θ:

$$\mathbb{E}_A\, log(\prod_{r=1}^{R} p(A^r, T^r|S^r)) = \mathbb{E}_A\, log(\prod_{r=1}^{R}\prod_{i=1}^{m_r}\frac{1}{n_r}\theta(t_i^r|s_{a_i^r}^r) = \mathbb{E}_A \sum_{r=1}^{R}\sum_{i=1}^{m_r}[log\frac{1}{n_r}+$$

$$+log\theta(t_i^r|s_{a_i^r}^r)] = -\sum_{r=1}^{R} m_r\, log\, n_r + \sum_{r=1}^{R}\sum_{i=1}^{m_r}\mathbb{E}_A\, log\,\theta(t_i^r|s_{a_i^r}^r) \quad (=)$$

Обозначим $C = -\sum\limits_{r=1}^{R} m_r\, log\, n_r$. Тогда получим:

$$(=)\;\; C + \sum_{r=1}^{R}\sum_{i=1}^{m_r}\sum_{j=1}^{n_r} p(a_i = j \mid T^r, S^r)\, log\,\theta(t_i^r|s_{a_i^r}^r)$$

Тогда получим следующую оптимизационную задачу:

$$\begin{cases} C + \sum\limits_{x,y} count(x,y)\, log\,\theta(y|x) \to \max\limits_{\theta} \\ \sum\limits_{y}\theta(y|x) = 1, \quad \forall x \in V^s \end{cases}$$

Где $count(x,y) = \sum\limits_{r=1}^{R}\sum\limits_{i=1}^{m_r}\sum\limits_{j=1}^{n_r} \mathbb{1}\{t_i^r = x\}\, \mathbb{1}\{s_j^r = y\}\, p(a_i = j \mid T^r, S^r)$, $\sum\limits_{x,y}$ - сумма по всевозможным парам (x,y), где $x \in V^s$ - произвольное слово из словаря исходного языка, а $y \in V^t$ из целевого.

Будем искать максимум через лагранжиан системы. Запишем его:

$$L = C + \sum_{x,y} count(x,y)\, log\,\theta(y|x) - \sum_{x}\lambda_x(\sum_{y}\theta(y|x) - 1)$$

Продифференцируем и приравняем к 0 частные производные:

$$\frac{\partial L}{\partial\theta(y|x)} = \frac{count(x,y)}{\theta(y|x)} - \lambda_x = 0 \quad\Rightarrow\quad \theta(y|x) = \frac{count(x,y)}{\lambda_x}$$

$$\frac{\partial L}{\partial \lambda_x} = \sum_y \theta(y|x) - 1 = 0 \;\;\Rightarrow\;\; \sum_y \theta(y|x) = 1$$

Просуммировав первое уравнение по всем $y \in V^t$, учитывая второе, получим:

$$1 = \sum_y \theta(y|x) = \frac{\sum_y count(x,y)}{\lambda_x} \;\;\Rightarrow\;\; \lambda_x = \sum_y count(x,y)$$

Таким образом, получим аналитический максимум по параметрам нашей модели:

$$\theta(y|x) = \frac{count(x,y)}{\sum_y count(x,y)}$$

References

Word Alignment and the Expectation-Maximization Algorithm. Adam Lopez. Johns Hopkins University.

Appendix

* Докажем, что знак суммы и произведения можно поменять местами в знаменателе. Для числителя всё аналогично.
$\sum\limits_{a_1=1}^{n} ... \sum\limits_{a_m=1}^{n} \prod\limits_{k=1}^{m} \theta(t_k|s_{a_k}) = \theta(t_1|s_1) \cdots \theta(t_m|s_1) + \theta(t_1|s_1) \cdots \theta(t_{m-1}|s_1)\theta(t_m|s_2) + ... + \theta(t_1|s_n) \cdots \theta(t_m|s_n) = [\theta(t_1|s_1) + ... + \theta(t_1|s_n)][\sum\limits_{a_2=1}^{n} ... \sum\limits_{a_m=1}^{n} \prod\limits_{k=2}^{m} \theta(t_k|s_{a_k})] = [\sum\limits_{a_1=1}^{n} \theta(t_1|s_{a_1})] \cdot [\sum\limits_{a_2=1}^{n} ... \sum\limits_{a_m=1}^{n} \prod\limits_{k=2}^{m} \theta(t_k|s_{a_k})] =$ {Аналгочино проделывая такие шаги по всем t_i, получим} $= [\sum\limits_{a_1=1}^{n} \theta(t_1|s_{a_1})] \cdots [\sum\limits_{a_m=1}^{n} \theta(t_m|s_{a_m})] = \prod\limits_{k=1}^{m} \sum\limits_{a_1=1}^{n} ... \sum\limits_{a_m=1}^{n} \theta(t_k|s_{a_k})$