8. IBM Model-1 の式の説明

内山将夫@NICT
mutiyama@nict.go.jp

IBM-Mode1の式の説明のまえに

Q:
何のために確率モデルを利用するか
A:

- 単語対応の確率や
- 単語対応自体を求めるため

単語対応の確率：
「お金」は「cash」と「money」のどちらに訳されることが多いか？
単語対応自体を求めるため：
「今日」「は」「良い」「天気」「だ」と「It」「is」「fine」「today」では,「今日」はどの単語に対応するか？

単語対応自体を求めるにはどうするか

条件：
英文$\mathbf{e}$と仏文$\mathbf{f}$が与えられていて，単語対応$\mathbf{a}$について，$P(\mathbf{f}, \mathbf{a}|\mathbf{e})$が計算できるとする
求めるもの$\hat{\mathbf{a}}$:

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}|\mathbf{e})$$

$\hat{\mathbf{a}}$は，$P(\mathbf{f}, \mathbf{a}|\mathbf{e})$を最大化するアラインメントである．これは最も確率が高いアラインメントである．
$\rightarrow$ アラインメントの確率をモデル化することにより，そのモデルの下での最適アラインメントを求めることができる．

単語対応の確率を求めにはどうするか

- 求めたいパラメタをθとして，明示的に式に導入すると

$$P(\mathbf{f}|\mathbf{e},\theta)$$

により，θをパラメタとしたときの$\mathbf{f}$の確率がわかる

- パラレルコーパスを$\mathbf{d}=\{\langle\mathbf{f},\mathbf{e}\rangle\}$とすると

$$L(\theta|\mathbf{d})=\log\prod_{\langle\mathbf{f},\mathbf{e}\rangle\in\mathbf{d}}P(\mathbf{f}|\mathbf{e},\theta)$$

によりパラメタθの下での，コーパス全体での$\mathbf{f}$の確率がわかる．このとき

$$\hat{\theta}=\arg\max_{\theta}L(\theta|\mathbf{d})$$

なる$\hat{\theta}$をみつけると，コーパスを最尤で生成するθが求まる．

ここまでのまとめ

確率でモデル化することにより

- 単語対応(アラインメント)が計算できる
- 単語対応確率等のパラメタが計算できる

→ 大きな利点

難しいところ

- モデルの正当性の評価が難しい
- 数式で色々なことを表現しないといけないので，あまり複雑なことは表現できない
- モデルの良さは，実際にデータに適用してみないとわからない
- 良いアイディアと思っても上手くいくかはわからない

IBM model-1の式展開

$$P(\mathbf{f},\mathbf{a}|\mathbf{e}) = P(m|\mathbf{e}) \prod_{j=1}^{m} P(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) P(f_j|a_1^{j}, f_1^{j-1}, m, \mathbf{e}) \quad (1)$$

$$\begin{aligned} P(m|\mathbf{e}) &= \text{仏文の長さ(単語数)が}m\text{である確率} \\ &= \text{ある定数}\epsilon \end{aligned}$$

$$\begin{aligned} P(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) &= j\text{番目の仏単語がつながるのが} \\ &\quad a_j\text{番目の英単語の確率} \\ &= \text{どの場所にも同確率} \\ &= \frac{1}{l+1} \end{aligned}$$

$$\begin{aligned} &P(f_j|a_1^{j}, f_1^{j-1}, m, \mathbf{e}) \\ &= j\text{番目の仏単語が}f_j\text{の確率} \\ &= f_j\text{と}e_{a_j}\text{のみで決まる} \\ &= t(f_j|e_{a_j}) \end{aligned}$$

以上より

$$P(\mathbf{f},\mathbf{a}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(f_j|e_{a_j}) \quad (2)$$

アラインメント$\mathbf{a}$について和をとる

$$
\begin{aligned}
& P(\mathbf{f}|\mathbf{e}) \\
&= \sum_{\mathbf{a}} P(\mathbf{f},\mathbf{a}|\mathbf{e}) \\
&= \sum_{\mathbf{a}} \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(f_j|e_{a_j}) \\
&= \frac{\epsilon}{(l+1)^m} \sum_{\mathbf{a}} \prod_{j=1}^{m} t(f_j|e_{a_j}) \\
&= \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) \qquad (3)
\end{aligned}
$$

$\mathbf{a} = a_1 ... a_m$におけるa_jは仏語f_jが英単語e_{a_j}に継がることを示す．また，$\mathbf{e} = e_1 ... e_l$で，かつ，$e_0 = \mathrm{NULL}$である． そのため，$0 \leq a_j \leq l$である．更に，$1 \leq j \leq m$なので，上式の ように変形される．

パラメタ推定

3 式において推定すべきパラメタは $t(f|e)$ のみである．これには前述のように，

$$\max P(\mathbf{f}|\mathbf{e})$$

となる $t(\cdot)$ を推定すれば良い．
制約として，各英単語 e について

$$\sum_f t(f|e) = 1$$

である．つまり，各英単語 e について，それに対応する f の確率の和は 1 である．

ラグランジェの未定乗数法により極値を求める

3式と制約より

$$h(t,\lambda) = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) - \sum_e \lambda_e (\sum_f t(f|e) - 1) \tag{4}$$

第2項が制約部分であり，これはλ_eで偏微分をすることにより，

$$\frac{\partial h}{\partial \lambda_e} = -(\sum_f t(f|e) - 1) = 0 \tag{5}$$

となる．つまり，λ_eによる偏微分の結果が0であることにより，極値におけるパラメタ値が制約を満すようになる．

一方，$t(f|e)$による偏微分においては，4式の第2項は，

$$\frac{\partial}{\partial t(f|e)} - \sum_e \lambda_e (\sum_f t(f|e) - 1) = -\lambda_e \tag{6}$$

4式の第1項は，まず，

$$\prod_{j=1}^{m} t(f_j|e_{a_j})$$

を偏微分する．

$$
\begin{aligned}
&\frac{\partial}{\partial t(f|e)} \prod_{j=1}^{m} t(f_j|e_{a_j}) \\
&= \left(\prod_{\substack{1\le k\le m \\ f_k\neq f \vee e_{a_k}\neq e}} t(f_k|e_{a_k}) \right) \frac{\partial}{\partial t(f|e)} t(f|e)^{\sum_{j=1}^{m}\delta(f,f_j)\delta(e,e_{a_j})} \\
&= \left(\prod_{\substack{1\le k\le m \\ f_k\neq f \vee e_{a_k}\neq e}} t(f_k|e_{a_k}) \right) n(e,f) t(f|e)^{n(e,f)-1} \\
&= n(e,f) t(f|e)^{-1} \left(\prod_{\substack{1\le k\le m \\ f_k\neq f \vee e_{a_k}\neq e}} t(f_k|e_{a_k}) \right) t(f|e)^{n(e,f)} \\
&= n(e,f) t(f|e)^{-1} \prod_{1\le k\le m} t(f_k|e_{a_k}) \qquad (7)
\end{aligned}
$$

ただし，

$$
\begin{aligned}
n(e,f) &= \sum_{j=1}^{m} \delta(f,f_j)\delta(e,e_{a_j}) \\
&= \text{仏単語が}\, f_j \,\text{で英単語が}\, a_j \,\text{の単語対応の数}
\end{aligned}
$$

- 1行目では，$t(f|e)$の項を取り出す
- 2行目では，$t(f|e)$で偏微分する
- 3,4 行目では，$t(f|e)^{-1}$ を括り出すことにより，$\Pi_{1\le k\le m}\, t(f_k|e_{a_k})$を再導入する

以上より，

$$\frac{\partial}{\partial t(f|e)}h(t,\lambda)$$
$$= \frac{\epsilon}{(l+1)^m}\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\left(\sum_{j=1}^{m}\delta(f,f_j)\delta(e,e_{a_j})\right)t(f|e)^{-1}$$
$$\prod_{1\leq k\leq m}t(f_k|e_{a_k})-\lambda_e \tag{8}$$

これを0とおくと

$$t(f|e) = \lambda_e^{-1}\frac{\epsilon}{(l+1)^m}$$
$$\times \sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\left(\sum_{j=1}^{m}\delta(f,f_j)\delta(e,e_{a_j})\right)\prod_{1\leq k\leq m}t(f_k|e_{a_k}) \tag{9}$$

まず，λ_eについて説明すると，これは4式において，$\Sigma_f\, t(f|e) = 1$となるように導入した変数である．この制約を満すには，$t(f|e) = \lambda_e^{-1}A(f|e)$としたとき，$\lambda_e = \Sigma_f\, A(f|e)$とすると，$\Sigma_f\, t(f|e) = 1$となる．つまり，$\lambda_e$は，単なる正規化定数である．

これまでのまとめ

- 9式により，$t(f|e)$ を表現した
- この式を利用すると $t(f|e)$ が求まると思われる
- しかし右辺にも $t(\cdot)$ がでてくる

EM 法により $t(\cdot)$ を求める

1. まず $t(\cdot)$ に適当な初期値を定める
2. 次に，その値を利用して，9式により $t(\cdot)$ を計算する
3. 2 を繰り返す

このようにして，$t(e|f)$ を求める道筋が整った．次は，9式を簡単化していく．

9式の再掲

$$t(f|e) = \lambda_e^{-1} \frac{\epsilon}{(l+1)^m} \times \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \left(\sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j}) \right) \prod_{1 \le k \le m} t(f_k|e_{a_k})$$

3式の再掲

$$P(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j})$$

これらより

$$t(f|e) = \lambda_e^{-1} \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}|\mathbf{e}) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j}) \qquad (10)$$

これを更に概念的に簡単化するために,

$$C(f|e; \mathbf{f}, \mathbf{e}) = \sum_{\mathbf{a}} P(\mathbf{a}|\mathbf{f}, \mathbf{e}) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j}) \qquad (11)$$

を定義する.

この式の$\Sigma_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j})$は，アラインメント$\mathbf{a}$において$f$と$e$が継がっている回数である．また，$P(\mathbf{a}|\mathbf{f}, \mathbf{e})$は，アラインメント$\mathbf{a}$の確率である．よって，$C(f|e; \mathbf{f}, \mathbf{e})$は，$f$と$e$が対応する回数の期待値である.

これから確認するように，$t(f|e)$は，fとeの対応の期待度数$C(f|e, \mathbf{f}, \mathbf{e})$を利用した割合により表現される.

$$
\begin{aligned}
t(f|e) &= \lambda_e^{-1} \sum_{\mathbf{a}} P(\mathbf{f},\mathbf{a}|\mathbf{e}) \sum_{j=1}^{m} \delta(f,f_j)\delta(e,e_{a_j}) \\
&= \lambda_e^{-1} \sum_{\mathbf{a}} P(\mathbf{f}|\mathbf{e}) P(\mathbf{a}|\mathbf{f},\mathbf{e}) \sum_{j=1}^{m} \delta(f,f_j)\delta(e,e_{a_j}) \\
&= \lambda_e^{-1} P(\mathbf{f}|\mathbf{e}) \sum_{\mathbf{a}} P(\mathbf{a}|\mathbf{f},\mathbf{e}) \sum_{j=1}^{m} \delta(f,f_j)\delta(e,e_{a_j}) \\
&= \lambda_e^{-1} P(\mathbf{f}|\mathbf{e}) C(f|e;\mathbf{f},\mathbf{e}) \\
&= \lambda_e'^{-1} C(f|e;\mathbf{f},\mathbf{e}) \qquad (12)
\end{aligned}
$$

$\lambda'_e = \lambda_e P(\mathbf{f}|\mathbf{e})^{-1}$は正規化定数．$\lambda'_e = \Sigma_f\, C(f|e;\mathbf{f},\mathbf{e})$とすれば，$\Sigma_f\, t(f|e) = 1$となる．

上記は，英文$\mathbf{e}$と仏文$\mathbf{f}$が一文だけしかない場合である．コーパス全体では，s番目の対訳文を$\mathbf{e}^{(s)}$と$\mathbf{f}^{(s)}$で表すと

$$
t(f|e) = \lambda_e'^{-1} \sum_{s} C(f|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)})
$$

である．これは，期待度数をコーパス全体で測定している．

これまでのまとめ

$$t(f|e) = \lambda'^{-1}_e \sum_s C(f|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)}) \quad (13)$$

$$\lambda'_e = \sum_f \sum_s C(f|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)}) \quad (14)$$

$$C(f|e;\mathbf{f},\mathbf{e}) = \sum_{\mathbf{a}} P(\mathbf{a}|\mathbf{f},\mathbf{e}) \sum_{j=1}^{m} \delta(f,f_j)\delta(e,e_{a_j}) \quad (15)$$

上式を利用した推定アルゴリズム

1. 全ての f と e について，$t(f|e)$ の初期値を選ぶ.
2. コーパス中の文 s について $C(f|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)})$ を計算する.
3. 各 e について，λ'_e を計算する
4. 各 f について，$t(f|e)$ を計算する
5. goto 2 or exit

残された問題

$$
\begin{aligned}
& C(f|e;\mathbf{f},\mathbf{e}) \\
&= \sum_{\mathbf{a}} P(\mathbf{a}|\mathbf{f},\mathbf{e}) \sum_{j=1}^{m} \delta(f,f_j)\delta(e,e_{a_j}) \\
&= \lambda_e P(\mathbf{f}|\mathbf{e})^{-1} t(f|e) \qquad (\text{式}12\text{参照}) \qquad (16)
\end{aligned}
$$

をどう計算するか？
4式に戻ると

$$
h(t,\lambda) = \frac{\epsilon}{(l+1)^m} \underline{\sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j})} \\
- \sum_e \lambda_e (\sum_f t(f|e) - 1)
$$

下線の部分で$(l+1)^m m$回の計算が必要になる．この部分を簡単化する必要がある．

$$\sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) = \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_i)$$

$(l+1)^m m$回の計算量が$m(l+1)$に減少した.
例により，この変形の正しさを確認する.
$m=3, l=1$のとき，$t(f_j|e_i) = t_{ji}$として,

$$\begin{aligned} \text{左辺} &= \sum_{a_1=0}^{1} \sum_{a_2=0}^{1} \sum_{a_3=0}^{1} \prod_{j=1}^{3} t(f_j|e_{a_j}) \\ &= t_{10}t_{20}t_{30} + t_{11}t_{20}t_{30} + \cdots + t_{11}t_{21}t_{30} + t_{11}t_{21}t_{31} \end{aligned}$$

$$\begin{aligned} \text{右辺} &= \prod_{j=1}^{3} \sum_{i=0}^{1} t(f_j|e_i) \\ &= (t_{10}+t_{11})(t_{20}+t_{21})(t_{30}+t_{31}) \end{aligned}$$

4式を簡単化する

$$h(t,\lambda) \;=\; \frac{\epsilon}{(l+1)^m}\prod_{j=1}^{m}\sum_{i=0}^{l} t(f_j|e_i) - \sum_e \lambda_e(\sum_f t(f|e) - 1)$$

第2項の$t(f|e)$による偏微分は$-\lambda_e$.

$$n_f \;=\; \sum_{j=1}^{m} \delta(f, f_j) = (f = f_j)\text{なる}\, f_j\text{の数}$$

$$n_e \;=\; \sum_{i=0}^{l} \delta(e, e_i) = (e = e_i)\text{なる}\, e_i\text{の数}$$

とすると

$$\begin{aligned}
&\prod_{j=1}^{m}\sum_{i=0}^{l} t(f_j|e_i) \\
&= \left(\prod_{\substack{1\le j\le m\\ f_j\neq f}}\sum_{i=0}^{l} t(f_j|e_i)\right)\left(\sum_{i=0}^{l} t(f|e_i)\right)^{n_f} \\
&= \left(\prod_{\substack{1\le j\le m\\ f_j\neq f}}\sum_{i=0}^{l} t(f_j|e_i)\right)\left(\sum_{\substack{0\le i\le l\\ e_i\neq e}} t(f|e_i) + n_e t(f|e)\right)^{n_f} \qquad (17)
\end{aligned}$$

だから

$$\begin{aligned}
&\frac{\partial\text{第}2\text{項}}{\partial t(f|e)} \\
&= n_f n_e \left(\sum_{\substack{0\le i\le l\\ e_i\neq e}} t(f|e_i) + n_e t(f|e)\right)^{n_f-1} \\
&= n_f n_e \left(\sum_{0\le i\le l} t(f|e_i)\right)^{n_f-1} \\
&= \frac{n_f n_e}{\Sigma_{0\le i\le l}\, t(f|e_i)} \left(\sum_{0\le i\le l} t(f|e_i)\right)^{n_f} \qquad (18)
\end{aligned}$$

以上より，

$$\frac{\partial}{\partial t(f|e)} \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_i) = \frac{n_f n_e}{\Sigma_{0\leq i\leq l}\, t(f|e_i)} \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_i) \tag{19}$$

よって，

$$\begin{aligned} &\frac{\partial h(t,\lambda)}{\partial t(f|e)} \\ &= \frac{n_f n_e}{\Sigma_{i=0}^{l}\, t(f|e_i)} \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_i) - \lambda_e \\ &= \frac{n_f n_e}{\Sigma_{i=0}^{l}\, t(f|e_i)} P(\mathbf{f}|\mathbf{e}) - \lambda_e \end{aligned} \tag{20}$$

よって，極値では $\frac{\partial h(t,\lambda)}{\partial t(f|e)} = 0$ より

$$\lambda_e P(\mathbf{f}|\mathbf{e})^{-1} = \frac{n_f n_e}{\Sigma_{i=0}^{l}\, t(f|e_i)}$$

よって，16式より

$$\begin{aligned} &C(f|e;\mathbf{f},\mathbf{e}) \\ &= \lambda_e P(\mathbf{f}|\mathbf{e})^{-1} t(f|e) \\ &= \frac{t(f|e)}{\Sigma_{i=0}^{l}\, t(f|e_i)} n_f n_e \end{aligned} \tag{21}$$

ここで，$\frac{t(f|e)}{\Sigma_{i=0}^{l}\, t(f|e_i)}$ は，f と e という対応に与えられる重みであり，$n_f n_e$ は，そのような対応の数である．

核となる式

$$C(f|e;\mathbf{f},\mathbf{e}) = \frac{t(f|e)}{\sum_{i=0}^{l} t(f|e_i)} n_f n_e \tag{22}$$

$$\lambda_e = \sum_f \sum_s C(f|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)}) \tag{23}$$

$$t(f|e) = \lambda_e^{-1} \sum_s C(f|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)}) \tag{24}$$

$t(\cdot)$推定のアルゴリズム

1. $t(f|e)$の初期値を設定する
2. 各文sについて$C(f|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)})$を計算する．
3. 各eについて，λ_eを計算する．
4. 各fについて，$t(f|e)$を計算する．
5. goto 2 or exit.

課題

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. (1993) The Mathematics of Statistical Machine Translation: Parameter Estimation.
を読んでみる．