19. 最小誤り率訓練

内山将夫@NICT
mutiyama@nict.go.jp

SMTの構成要素

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \sum_i \lambda_i h_i(\mathbf{e}, \mathbf{f})$$

- 探索： $\arg\max_{\mathbf{e}}$ なる $\hat{\mathbf{e}}$ の探索
- モデリング： 良い素性 $h_i(\mathbf{e}, \mathbf{f})$ の設計
- パラメタ調整：λ_i の学習

最小誤り率訓練 (MERT, Minimum Error Rate Training) は，パラメタ調整に利用される．

パラメタ調整の枠組

- 訓練データ： $h_i(\mathbf{e}, \mathbf{f})$ を獲得する
- 開発データ： λ_i を獲得する
- テストデータ： 翻訳性能を測定する

パラメタ調整の原則

- 翻訳性能を最大化するパラメタが欲しい
  翻訳性能をBLEUで測定するとすると，BLEUを最大化するようなパラメタが欲しい.

開発データにおける入力文を

$$F = \{\mathbf{f}_1, \mathbf{f}_2, ...\}$$

参照用の翻訳文を

$$R = \{\mathbf{r}_1, \mathbf{r}_2, ...\}$$

F を機械翻訳した結果を

$$E = \{\mathbf{e}_1, \mathbf{e}_2, ...\}$$

としたとき,

$$\hat{\lambda} = \arg\max_{\lambda} \mathrm{BLEU}(R, E)$$

なるパラメタ $\hat{\lambda}$ が欲しい.

最適化としての$\hat{\lambda}$の探索

1. $\lambda_m =$ 適当な初期値, $C_i = \phi$
2. for $\mathbf{f}_i \in F$
   (a) $C'_i = \{\mathbf{e}_{i,s} |$ スコア$\Sigma \lambda_m h_m(\mathbf{e}, \mathbf{f}_i)$が大きい$n$翻訳文$\}$
   (b) $C_i = C_i \cup C'_i$. (これまでの翻訳候補に加えて，今のλを利用して得られた翻訳候補を追加する)
3. λを更新する
   (a) 今のλを利用して，拡張されたC_iの中から一番スコアが高い$\mathbf{e}_i = \arg\max_{\mathbf{e}} \Sigma lambda_m h_m(\mathbf{e}, \mathbf{f}_i)$なる$\mathbf{e}_i$を得ることにより，$\mathbf{f}_i$に対する，今のパラメタでの翻訳文とする．
   (b) $E = \{\mathbf{e}_i |$ 上記で選ばれた翻訳文$\}$を利用して，λに対応する$\mathrm{BLEU}(E, R; \lambda)$を得る．
   (c) これにより，$\lambda \rightarrow \mathrm{BLEU}(E, R; \lambda)$ の関係が計算できるので，λ を少しずつ変えながら，現在の翻訳文集合C_iから，なるべくBLEU が大きくなるように，$\mathbf{e}_i$を選択できるようなλを探す
4. goto 2 or exit

多変量最適化の方法

- Simplex 法，Powell 法等のノンパラメトリック法(関数勾配が不要な方法)

  cf. Numerical Recipes in C

- 対数線形モデルに特有な方法

対数線形モデルに特有な方法

- ある方向 $\mathbf{d}$ について，1次元最適化をする
- 上記を，たくさんの方向に繰り返して，少しずつ解を改善して，最適解を求める．

1次元最適化を高速化する．

最小誤り率訓練

BLEU 最大化の代りに，より簡単な，誤り個数最小化の問題を考える．これを，あとで，BLEU 最大化に拡張する

誤り個数 $E(\mathbf{r}_1^S, \mathbf{e}_1^S)$ の定義

$$E(\mathbf{r}_1^s, \mathbf{e}_1^s) = \sum_{s=1}^{S} E(\mathbf{r}_s, \mathbf{e}_s)$$

$$\text{参照文のリスト} = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_S\}$$

$$\text{翻訳文のリスト} = \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_S\}$$

$$E(\mathbf{r}_s, \mathbf{e}_s) = \begin{cases} 1 & (\mathbf{r}_s \neq \mathbf{e}_s) \\ 0 & (\mathbf{r}_s = \mathbf{e}_s) \end{cases}$$

この $E(\mathbf{r}_s, \mathbf{e}_s)$ が，文が完全に一致するときに 0 で，そうでないときに 1 となっているので，翻訳文の評価としては，ずいぶんと簡略化されている．

誤り個数を最小化するパラメタ $\hat{\lambda}$

$$\hat{\lambda}_1^M = \arg\min_{\lambda_1^M} \sum_{s=1}^{S} E(\mathbf{r}_s, \mathbf{e}(\mathbf{f}_s; \lambda_1^M))$$

$$\mathbf{e}_s = \mathbf{e}(\mathbf{f}_s; \lambda_1^M) = \arg\max_{\mathbf{e} \in C_s} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}, \mathbf{f}_s)$$

$$\begin{aligned} C_s &= n\text{個の翻訳候補} \\ \lambda_m &= \text{素性}\, m\, \text{の重み} \\ h_m(\mathbf{e}, \mathbf{f}_s) &= \text{素性}\, m\, \text{の値} \\ \mathbf{f}_s &= \text{入力文} \end{aligned}$$

誤り個数の性質

$$E(\mathbf{r}_1^s, \mathbf{e}_1^s) = \sum_{s=1}^{S} E(\mathbf{r}_s, \mathbf{e}_s)$$

- $E(\mathbf{r}_s, \mathbf{e}_s)$ の和が全体の値となる
- したがって，個々の誤り $E(\mathbf{r}_s, \mathbf{e}_s)$ と λ_1^M の関係がわかれば，それを加算すれば，全体の誤りと λ_1^M の関係がわかる．

さて，一次元の最適化では，ある方向 $\mathbf{d}$ に向けての最適化をする．その方向をM次元ベクトル $\mathbf{d}_1^M$ により表現する．すると，ある定数ベクトル $\mathbf{g}_1^M$ を利用することにより，素性ベクトル λ_1^M は

$$\lambda_1^M = \mathbf{g}_1^M + \gamma \mathbf{d}_1^M$$

と表現できる．
したがって，λ_1^M を，ある与えられた方向 $\mathbf{d}_1^M$ に最適化するとは，$E(\mathbf{r}_1^s, \mathbf{e}_1^s)$ が最小となるような，$\mathbf{g}_1^M$ と γ を求めることである．
ここで，

$$\mathbf{e}_s = \mathbf{e}(\mathbf{f}_s; \lambda_1^M) = \arg\max_{\mathbf{e} \in C_s} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}, \mathbf{f}_s)$$

により，候補 C_s から $\mathbf{e}_s$ を選んで，それにより，$E(\mathbf{r}_s, \mathbf{e}_s)$ が決まる．この $\mathbf{e}_s$ が，λ_1^M，つまり，$\mathbf{g}_1^M$ と γ により異なる．
したがって，$\mathbf{e}_s$ と $\mathbf{g}_1^M$，γ の関係が知りたい．

$\mathbf{e}_s$ と $\mathbf{g}_1^M$，γ の関係

素性値のベクトルを $\mathbf{h}_1^M = \{h_1(\mathbf{e}, \mathbf{f}), ...\}$ とする．すると，翻訳文集合 C_s 中の候補を $\mathbf{e}_i$ とすると，そのスコア s_i は,

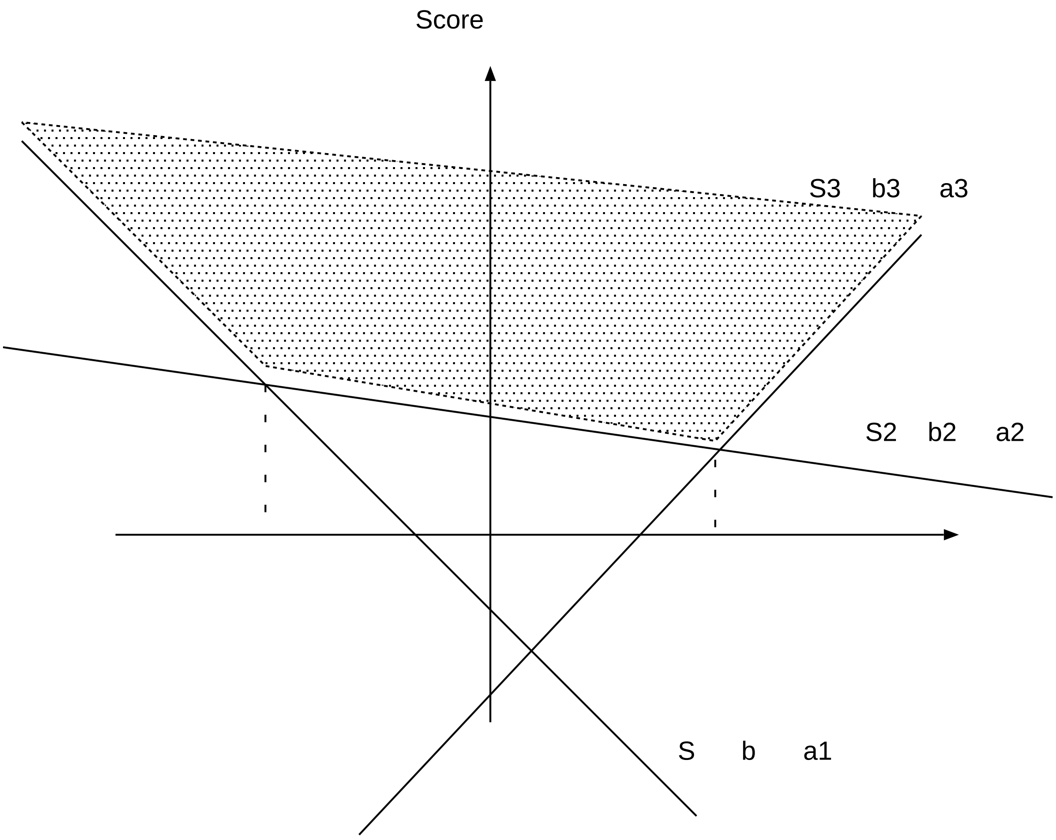
$$\begin{aligned} s_i &= \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}_i, \mathbf{f}_s) \\ &= \lambda_1^M \cdot \mathbf{h}_1^M \\ &= (\mathbf{g}_1^M + \gamma \mathbf{d}_1^M) \cdot \mathbf{h}_1^M \\ &= (\mathbf{g}_1^M \cdot \mathbf{h}_1^M + \gamma \mathbf{d}_1^M \cdot \mathbf{h}_1^M) \\ &= (b_i + \gamma a_i) \end{aligned}$$

ただし,

$$\begin{aligned} b_i &= \mathbf{g}_1^M \cdot \mathbf{h}_1^M \\ a_i &= \mathbf{d}_1^M \cdot \mathbf{h}_1^M \end{aligned} \tag{1}$$

である．つまり，スコア s_i は，ある定数 a_i と b_i から定められる直線 $a_i + \gamma b_i$ の上にある．すなわち，$\mathbf{e}_i$ のスコア s_i は，γ を変えると変わる．

γ の変化による $\mathbf{e}_s = \arg\max_{\mathbf{e}_i} s_i$ の変化



- $(-\infty, \gamma_1]$ のときは スコア S_1 が最大なので文 e_1 が選ばれる.
- $(-\gamma_1, \gamma_2]$ のときは，e_2 が選ばれる
- $(-\gamma_2, \infty]$ のときは，e_3 が選ばれる

初期値 $= E(\gamma = -\infty) = E(\mathbf{r}, \mathbf{e}_1)$
左から右に動いていって，
γ_1 になったら $\Delta E = E(\mathbf{r}, \mathbf{e}_2) - E(\mathbf{r}, \mathbf{e}_1)$
γ_2 になったら $\Delta E = E(\mathbf{r}, \mathbf{e}_3) - E(\mathbf{r}, \mathbf{e}_2)$
のように，γ の変化と，その時点での ΔE を記録する．
すると，ある参照文 $\mathbf{r}$ について，γ を動かしていったときに，どの時点で，誤りの個数が変化するかがわかる．

各入力文についての結果を統合する

- 入力文1
  $\gamma_1^1 \;\rightarrow\; \Delta E_1^1$
  $\gamma_2^1 \;\rightarrow\; \Delta E_2^1$
  $\ldots$
- 入力文2
  $\gamma_1^2 \;\rightarrow\; \Delta E_1^2$
  $\gamma_2^2 \;\rightarrow\; \Delta E_2^2$
  $\ldots$

これらをみんなあわせると
$\gamma_1 \;\rightarrow\; \Delta E_1$
$\gamma_2 \;\rightarrow\; \Delta E_2$
$\ldots$
のように，どのγにおいて，どの程度，誤りの個数が変化したかがわかる．
これより，$\gamma = -\infty$ のときの誤り個数に対して，$\gamma_1, \gamma_2,$ $\ldots$ とγを変えていったときの誤りの個数の変化がわかるので，そのときの最小誤りのところのγ を利用する．
このγを利用すると，一次元方向での最小化が達成できる．そのため，この結果を利用することにより，多次元の最適化ができる．

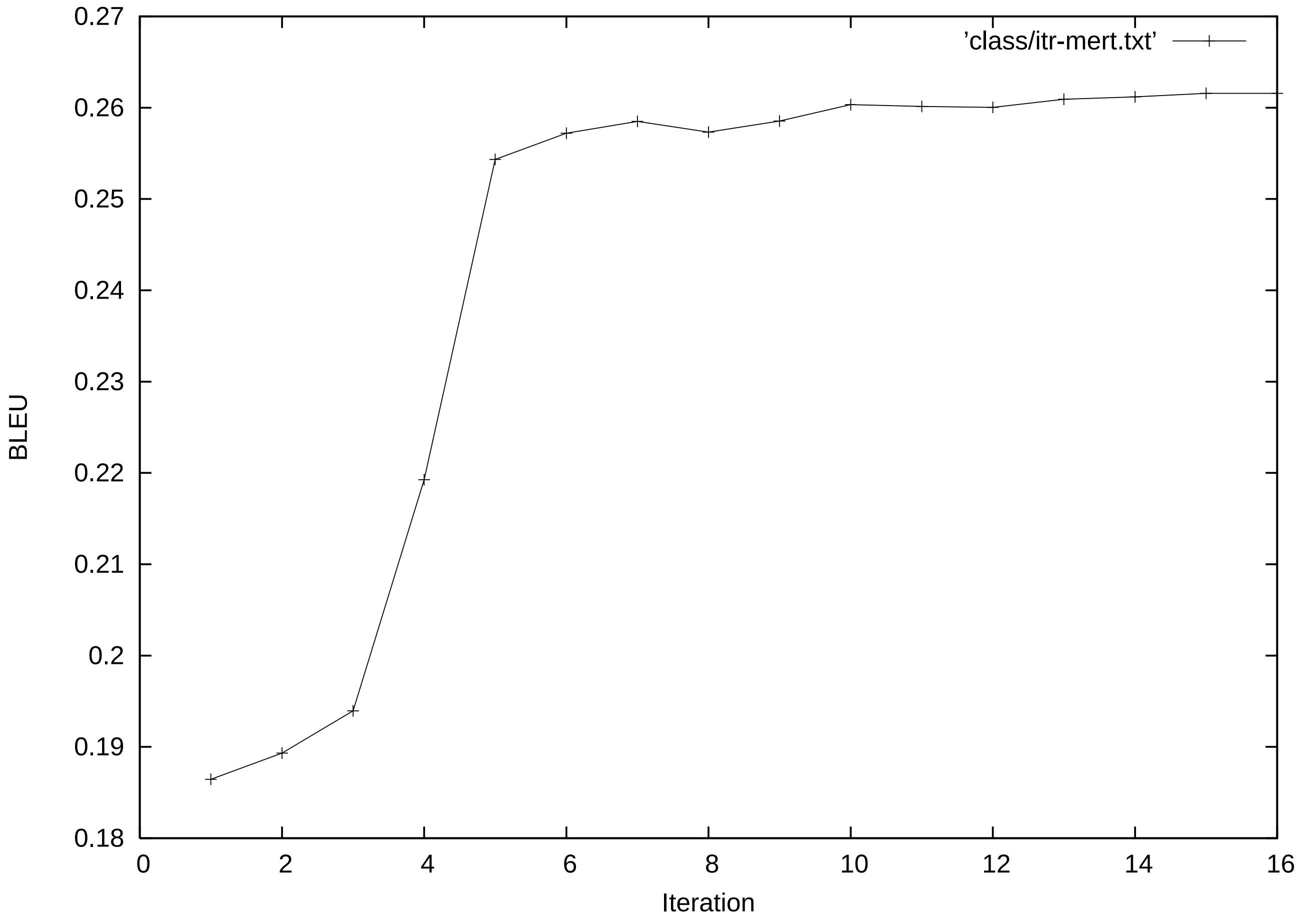
BLEUへの拡張

$$\mathrm{BLEU} = BP(\cdot)\exp(\sum_{n=1}^{N}\frac{\log p_n}{N}) \quad (2)$$

$$\begin{aligned}
BP(\cdot) &= \text{長さの短い文へのペナルティ} \\
N &= 4 \\
p_n &= \text{ngram精度} \\
&= \frac{\sum_{\text{MT訳}}\sum_{\text{MT訳のngram}}\text{共有ngram数}}{\sum_{\text{MT訳}}\sum_{\text{MT訳のngram}}\text{ngram数}}
\end{aligned}$$

拡張へのポイントは，BLEU においては，p_nのような部分が，共有するngramの各文に対する総和として表されることである．したがって，誤りの場合と同様にγが変化するたびに，共有するngram の総和が変化するので，そのたびに，共有するngramの総和等からp_n等を計算すれば，γが変化するたびのBLEUの変化がわかる．したがって，単純な誤り個数の場合と同様に，BLEUが最大となるγがわかる．

MERTの繰り返し回数とBLEUの関係



BLEUの変化が大きいことがわかる．パラメタ値の調整は，質の良い翻訳を達成するために，必要不可欠である．

繰り返し回数と訳文の変化1

Input: thus , the left input data of node nd15 is obtained .
Reference: これにより、ノードＮＤ１５の左入力データが得られたことになる。

itr1: したがって 、 左 入力 データ が ノード ＮＤ １５ が 得 られる 。
itr2: 以上のようにして構成されているがされているノードＮＤ１５のようにして得られたままに放置しているとされてい るのが、、、、のデータのデータのようにして、入力されている ように されている 。
itr3: このようにして、ノードＮＤ１５の 左 入力 データ が 得 られる ように なって いる ように なっ て いる 。
itr4: これ に より 、 ノード ＮＤ １５ の 左 入力 データ が 得 られる 。

繰り返し回数と訳文の変化2

Input: the system arrangement shown in fig. 1 will be described in more detail below .
Reference: 図 1 の システム 構成 について 更に 詳細 に 説明 する 。

itr1: この 構成 に 詳しく 説明 する 。
itr2: より 以 上 の よ う に 構 成 さ れ て い る の は 、 図 1 の 第 1 の 図 に 示 す よ う に し た 場 合 に つ い て 説 明 し た よ う に 構 成 さ れ て い る さ れ て い る シ ス テ ム の 詳 細 な 説 明 さ れ る よ う に なっ て い る の 下 方 に 位 置 し て 説 明 す る よ う に 構 成 さ れ て いる よう に さ れ て いる 。
itr3: 図 1 に 示 す よ う に 構 成 さ れ て 、 システム の より 詳細 に 説明 する 以下 の よ う に なっ て いる 。
itr4: 図 1 に 示す よう に 構成 さ れ 、 システム の 更に 詳細 に 説明 する 。
itr5: 図 1 に 示す システム 構成 の 詳しく 説明 する 。
itr6: 図 1 に 示す 構成 の システム 詳しく 説明 する 。
itr7: 図 1 に 示す システム 構成 の 更に 詳細 に 説明 する 。

まとめ

- BLEU が最大化するようにパラメタを調整することにより，評価値に沿ったパラメタを獲得できる

これより，適正な評価を与えることが重要であることがわかる．

- 自動的に適正な評価を与えることができれば，その評価を最大化するようにパラメタを調整することにより，良いシステムを作成することができる．