# 文脈特徴を用いたCRFによる音声認識誤り訂正*

☆中谷良平, 滝口哲也, 有木康雄 (神戸大)

## 1　はじめに

大語彙連続音声認識において，Conditional Random Fields (CRF) [1] を用いて認識結果中の誤りを検出する手法が提案されている [2]．CRF は，誤り部分を特徴づける不自然な N-gram だけでなく，品詞情報や信頼度など，様々な素性を自由に使って誤り傾向を学習できる．本稿では，この CRF を用いて音声認識誤り訂正を行うことを目的としている．素性としては，長距離言語情報などを用いる．提案手法を，日本語話し言葉コーパスに適用したところ，単語誤り率の改善が見られたので報告する．

## 2　モデル学習と誤り訂正

### 2.1　CRF による誤り検出モデルの学習

誤り検出モデルの学習では，学習に音声認識結果と，対応する正解文書を用い，正解部分，誤り部分で出現しやすい特徴を学習する．その結果，不自然な N-gram の発生を抑えることができる．また，表層単語の N-gram に限らず，文脈情報や認識スコアなど，様々な言語情報が誤り傾向学習に有効であると考えられる．本研究では誤り検出モデルを，認識結果に付与された複数の情報から，各単語に対して正解か誤りかのラベルを付与していく系列ラベリング問題と考え，CRF でモデル化する．

CRF では，入力記号列 $x$ に対する出力ラベル列 $y$ の条件付確率分布 $P(y \mid x)$ を次式のように定義する．

$$P(y \mid x) = \frac{1}{Z(x)} \exp(\sum_a \lambda_a f_a(y, x)) \tag{1}$$

ここで $f_a$ は素性，$\lambda_a$ は素性関数に対する重みとなる．$Z(x)$ は分配関数で，次式で与えられる．

$$Z(x) = \sum_y \exp(\sum_a \lambda_a f_a(y, x)) \tag{2}$$

パラメータ $\lambda_a$ は，学習データが与えられたとき，条件付確率分布 (1) の対数尤度を最大にするように学習される．

識別は学習によって得られた確率分布関数 $P(y \mid x)$ を用いて，与えられた入力記号列 $x$ に対する最適な出力ラベル列 $\hat{y}$ を求める問題となる．

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(y \mid x) \tag{3}$$

本稿では，このモデルを用いて競合候補から正解を選び出すことで誤り訂正を実現する．

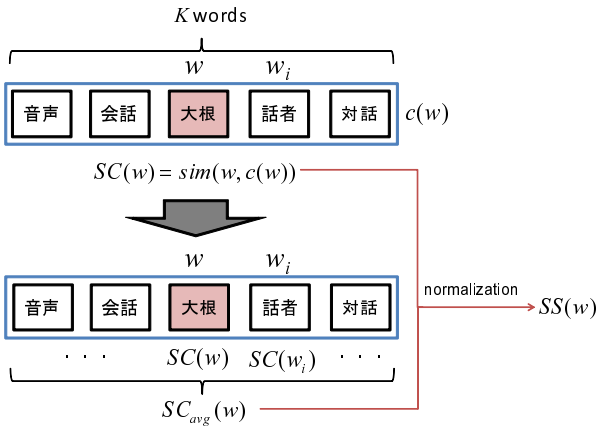


Fig. 1　意味スコアの算出

### 2.2　長距離言語情報

本稿では，文脈情報として意味スコアを用いる．意味スコアとは，周辺の認識結果単語を参照したときに，識別対象単語の出現が不自然でないかという情報のことである．例えば Fig. 1 のように,「音声」,「会話」,「話者」,「対話」などが含まれる単語列の中に,「大根」という単語が含まれる場合，明らかに不自然である．この存在単語の自然さを意味スコアとして算出し，誤り検出に用いる．しかし，意味スコアは，どの単語と共起しても不自然でない「は」や「です」といった機能語に対しては意味をなさない．そのため，本稿では内容語として名詞，動詞，形容詞のみに意味スコアを与える．内容語 $w$ の意味スコア $SS(w)$ の算出手順は次の通りである．

1. $w$ の周辺に現れる内容語を，Fig. 1 のように文脈窓幅 $K$ で集め，単語集合 $c(w)$ とする ($w$ 自身も含む)．
2. $c(w)$ 内の各単語 $w_i$ について，$c(w)$ 内の他の単語との類似度 $sim(w_i, c(w))$ を求め，$SC(w_i)$ とする．

$$SC(w_i) = sim(w_i, c(w)) \tag{4}$$

3. $SC(w_i)$ から，平均 $SC_{avg}(w)$ を求める．

$$SC_{avg}(w) = \frac{1}{K} \sum_i SC(w_i) \tag{5}$$

4. $SC(w)$ と $SC_{avg}(w)$ の差を意味スコア $SS(w)$ とする．

$$SS(w) = SC(w) - SC_{avg}(w) \tag{6}$$

*Speech Recognition Error Correction Using Context Features Based on CRF, by Ryohei Nakatani, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

Table 1　実験に用いたデータ

| | 学習 | 評価 |
|---|---|---|
| 講演数 | 150 | 13 |
| 発話数 | 39,808 | 4,771 |
| 単語数 | 361,513 | 39,822 |

Table 2　単語誤り率と誤り種類別の評価

| | SUB | DEL | INS | COR | WER |
|---|---|---|---|---|---|
| CN-oracle | 1,753 | 2,128 | 851 | 35,521 | 11.88 |
| CN-best | 7,040 | 1,794 | 3,447 | 30,568 | 30.84 |
| Nonsemantic | 6,359 | 2,275 | 2,262 | 30,768 | 27.36 |
| Proposed method | 6,257 | 2,344 | 2,193 | 30,801 | 27.11 |

ステップ 2 で出てくる単語間類似度 $sim(w_i, c(w))$ の算出には，潜在的意味解析 (Latent Semantic Analisys : LSA) [3] を用いた．LSA は大量のテキストにおける単語の共起関係を統計的に解析することで，学習データに直接の共起がない単語間の類似度についても求めることができる手法である．

## 3　評価実験

### 3.1　実験条件

音声認識器 Julius [4] の認識結果に対して，提案した誤り訂正の評価実験を行った．データは日本語話し言葉コーパス (CSJ) を用いた．音響モデル，言語モデルともに CSJ から学習した．LSA の学習には，CSJ の書き起こし文書のうち，評価データを含まない 2,672 講演を用いた．誤り訂正を行う際の競合候補には Confusion Network [5] を用いる．意味スコアを求める際の単語集合 $c(w)$ は，前後 3 発話ずつの Confusion Network における存在確率最大の単語列に，識別対象単語 $w$ を加えたものとした．CRF による誤り検出モデルの学習と評価実験に用いたデータは，Table 1 のようになっている．学習する素性は，表層単語 bigram，trigram，CN に付与されている存在確率，LSA による意味スコアである．

### 3.2　実験結果

実験結果を Table 2 に示す．「SUB」は置換誤り，「DEL」は削除誤り，「INS」は挿入誤りの数をそれぞれ表している．「COR」は正解単語の数，「WER」は単語誤り率である．「CN-oracle」は，Confusion Set において常に正解の単語を選択したときの WER である．ただし，正解がないときはその Confusion Set 中で最も存在確率の高い単語を選んでいるため，ヌル遷移が選択されることで削除誤りが最小にはなっていない．「CN-best」は，誤り訂正前のベースとなる Confusion Network の最尤候補列の単語誤り率で，「Proposed method」は，本研究の提案手法である．また「Nonsemantic」は，提案手法の素性として意味スコアを用いない場合の単語誤り率の改善を表している．

表より，Proposed，Nonsemantic ともにベースとなる CN-best 単語誤り率が改善している．意味スコアを使わない Nonsemantic でも CN-best に比べて WER が 3.48 ポイント低くなったことから，CRF によって適切に誤り訂正が行われていることがわかる．また、意味スコアを追加した提案手法では，さらに 0.25 ポイント改善し，CN-best と比較すると 3.73 ポイント改善した．

## 4　おわりに

本稿では，CRF による誤り検出を利用して，Confusion Network 上の誤りを訂正することで，音声認識精度の改善を行った．誤り傾向学習のために様々な素性を取り入れ，特に意味スコアを取り入れることによって，長距離言語情報を考慮した誤り訂正が可能になった．

今後の課題として，CRF による誤り検出精度の改善が考えられる．CRF で学習する際の素性として品詞情報，例えば品詞の trigram などを用いることも有効であると考えられる．その他に，CRF の改良手法 [6] による学習を取り入れることも考えたい．

## 参考文献


[1] J. Lafferty, *et al.* “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” ICML, pp. 282-289, 2001.

[2] 松本, 他, “複数の言語情報を用いた CRF による音声認識誤りの検出”, 音講論 (春), pp. 227-228, 2009.

[3] Jerome R. Bellegarda, “Latent semantic mapping,” IEEE Signal Processing, 5(22), pp. 70-80, 2005.

[4] “Julius,” http://julius.sourceforge.jp/

[5] L. Mangu, *et al.* “Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks,” Computer Speech and Language, pp. 373-400, 2000.

[6] Jian Peng, Liefeng Bo, Jinbo Xu, “Conditional Neural Fields,” *NIPS22*, pp. 1419-1427, 2009.