CRF と Confusion Network を用いた音声認識誤り訂正*

☆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，ニュースなどで読み上げられる書き言葉は，単語正解精度で 95%程度の認識が可能である [1]．また，学会講演音声のような話し言葉でも，85%程度の精度で認識できるようになってきた．しかし，まだ十分な音声認識精度が得られたわけではない．機械に誤認識した単語列の不自然さを学習させれば，もっと認識精度を改善できると考えられる．

そこで，複数の仮説をリランキングすることで認識精度を向上させる，認識誤り訂正手法が提案されている [2][3]．この手法は，音声認識器の出力した仮説と，それに対応する正解単語列から，認識誤りを特徴づける不自然な N-gram を学習し，リランキングに用いる．

また，文脈情報を用いた，CRF [4] による音声認識誤り検出手法が提案されている [5]．長距離文脈的に不自然な単語連鎖や，品詞の N-gram，活用形-品詞の連鎖などの文法的な知識を学習し，認識誤り検出精度が向上することが報告されている．

そこで本稿では，CRF を用いた誤り訂正手法を提案する．CRF を用いることにより，様々な素性を柔軟に取り込むことができる．また，CRF によって学習，訂正する仮説集合として，Confusion Network を導入する．CRF により誤りであると判定された単語について，この Confusion Network の候補の中から，正しい候補を選択することで誤り訂正を実現する．これによって，N-best や単語グラフを用いた誤り訂正と違い，単語単位での柔軟な誤り訂正が可能になる．この提案手法による，日本語話し言葉コーパスに対する実験から，単語認識率の改善が見られたので報告する．

2 長距離文脈情報を用いた CRF による誤り検出手法

2.1 長距離文脈情報

本稿で用いる長距離文脈情報とは，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．人間は，N-gram のような部分的な文脈情報だけでなく，より広範囲に渡る長距離文脈情報も考慮しながら音声を聞きとっていると考えられる．例えば Fig. 1 のように,「犯罪」,「裁判」,「無罪」,「弁護士」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．音声認識結果に出現した内容語 w の意味スコア $SS(w)$ の算出手順は次の通りである [5].

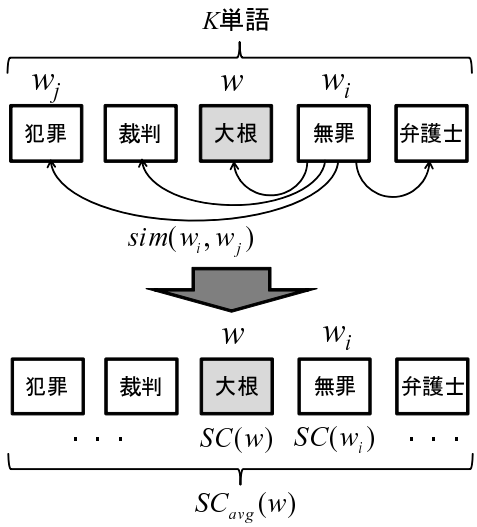

Fig. 1 意味スコアの算出

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 $M(w)$ とする (w 自身も含む).
2. $M(w)$ 内の各単語 w_i について，$M(w)$ 内の他の単語との類似度 $sim(w_i, w_j)$ を求め，その平均を $SC(w_i)$ とする.

$$SC(w_i) = \frac{1}{K} \sum_j sim(w_i, w_j). \quad (1)$$

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

$$SC_{avg}(w) = \frac{1}{K} \sum_i SC(w_i). \quad (2)$$

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

$$SS(w) = SC(w) - SC_{avg}(w). \quad (3)$$

*Speech Recognition Error Correction using CRF and Confusion Network, by Ryohei Nakatani, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

ステップ 2 で出てくる単語間類似度 $sim(w_i, w_j)$ の算出には，潜在的意味解析 (Latent Semantic Analisys : LSA) [6] 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．

2.2 CRF による誤り検出

誤り傾向学習では，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．例えば,「の-よう-は」のような不自然な N-gram が出現すれば,「は」は誤りである可能性が高いということが考えられる．また，表層単語の N-gram に限らず，意味スコアが低いと誤認識の可能性があるといった情報も有効である．本稿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，Conditional Random Field (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．

$$P(y \mid x) = \frac{1}{Z(x)} \exp(\sum_a \lambda_a f_a(y, x)). \quad (4)$$

ここで f_a は素性，λ_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．

$$Z(x) = \sum_y \exp(\sum_a \lambda_a f_a(y, x)). \quad (5)$$

パラメータ λ_a は，学習データ $(x_i, y_i)(1 \leq i \leq N)$ が与えられたとき，条件付確率分布 (4) の対数尤度，

$$\mathcal{L} = \sum_{i=1}^{N} \log P(y_i \mid x_i). \quad (6)$$

を最大にするように学習される．これは，正解ラベル列のコストと他のすべてのラベル列のコストとの差が最大になるように学習することに相当する．学習は，準ニュートン法である L-BFGS 法によって行われる．

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．$\hat{y}$ は次式をもとに Viterbi アルゴリズムにより効率的に求めることができる．

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(y \mid x). \quad (7)$$

3 Confusion Network 上での誤り訂正

3.1 Confusion Network

Confusion Network (CN) とは，音声認識器の内部状態を簡潔かつ高精度なネットワーク構造へ変換したもので，単語誤り最小化に基づいた音声認識における中間結果である．単語グラフを音響的なクラスタリングにより，リニアな形式に圧縮することで求められる [7]．CN は仮説の組み合わせにより，N-best 文リストよりも多くの候補を生成することができる．その結果，正しく正解を選び出した場合，日本語話し言葉コーパスにおいて，およそ 95%の正解を得ることができる可能性がある [8].

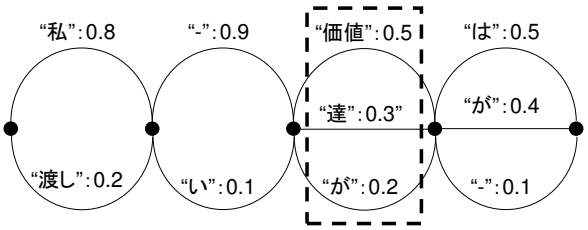

Fig. 2　Confusion Network の具体例

Fig. 2 は,“ 私達は ”という発話を入力したときの CN で，仮説集合のような形で表現される．破線で囲まれた遷移の集合は，時間的な競合候補を表しており，この集合を Confusion Set (CS) と呼ぶ.“ - ”は，その CS には単語が存在しないことを表現していて，ヌル遷移と呼ばれる．ヌル遷移を選択することでその CS をスキップすることができる．本稿では，この CS の中から正しい認識結果を選び出すことで誤り訂正を行う．

また，CN 上の各単語は，CS における存在確率 (信頼度) を持っている．CS はこの存在確率の高い順にソートされるので，第一候補が最も存在確率が高く，以降の競合候補は順に存在確率が低くなっていく．よって，Fig. 2 における第一候補（最上段）を選択していくと，最尤候補が得られる．

3.2 誤り訂正手法

本稿では前述したように CRF と Confusion Network を用いて誤り訂正を行う．CN を用いることで単語ごとに誤りを訂正できるため，より柔軟な誤り訂正が可能になる．

提案手法では，まず正解と誤りを CRF によってラベル付けする，誤り検出ための学習を行う．音声認識器が出力した認識結果 (Confusion Network) と正解文書を用いて，認識結果に正誤ラベリングを行う．ヌル遷移に対応するため，一般的な 1-best ではなく，CN の第一候補単語列（最尤候補）を CRF によって学習する．学習に用いる素性は，4 章の評価実験で述べる．学習後，以下のアルゴリズムに従って誤り訂正を行う．

1. 評価データを音声認識後，Confusion Network を出力する．

2. CN の第一候補列のみを抜き出し，CRF による誤り検出を行う．
3. 誤りと判定された語は，対応する CS から次の候補を選び出し，置き換えてもう一度誤り検出を行う．
4. CS の中に正解と思われる語が存在しなければ，存在確率の最も高い語を選択する．
5. すべての CS について順番に 3,4 を繰り返す．

このアルゴリズムの結果，CRF により誤りと判定された語が，正解と判定された語で訂正される．また，ステップ 5 で「順番に」と述べたのは，CRF によって学習する際の素性として bigram，trigram を用いていることから，前の単語が訂正されると，後ろの単語の正誤判定が変わることがあるためである．例えば，2 連続で誤りラベルが付けられている単語列について，1 つ目の単語が訂正されると，bigram 特徴から，2 つ目の単語も正解ラベルに変わることがある．

4 評価実験

4.1 実験条件

音声認識器 Julius [9] 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．認識結果の単語認識率を 30%から 80%まで，10%ずつ増やしていき，それぞれについて Confusion Network 上でどの程度誤りを訂正できるのか，評価している．指標としては，単語認識率を用いた．また，同様の実験を N-best に対しても行い，比較を行った．N-best は $N = 100$ とし，誤り傾向は 1-best で学習した．

次に実験条件について述べる．データは日本語話し言葉コーパス (CSJ) を用いた．音声認識の音響モデルには，CSJ の 953 講演（男性 787 講演 + 女性 166 講演）の音声から作成した HMM を用いた．1 状態あたりの混合分布数は 16 としている．サンプリング周波数は 16kHz，音響特徴量は 12 次元 MFCC と対数パワー，12 次元 MFCC の一次微分を加えた 25 次元である．言語モデルには，CSJ から学習した trigram を用いている．意味スコアを求める際の単語集合 $M(w)$ は，前後 2 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，その学習には，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．意味スコアは，内容語として名詞，動詞，形容詞に対してのみ学習している．

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．学習データとなる CN の第一候補列の各単語には，CRF で学習するための正誤のラベルが必要となるが，CN と正解文書との間で DP マッチングをすることで自動でラベリングを行った．

Table 1　実験に用いたデータ

	学習	評価
講演数	150	16
発話数	39,808	5,893
単語数	367,715	48,558

4.2 実験結果

Table 2 は，認識結果の単語認識率を少しずつ変化させて，N-best と Confusion Network で，それぞれどの程度の認識率の改善が見られるかを表している．N-best を用いた手法，CN を用いた手法のいずれもベースの単語認識率から改善がみられた．それぞれについて比較すると，ベースとなる認識率が 30%のとき，N-best が CN を上回っているが，ベースが増加するにつれてその差が小さくなり，認識率が 60%になるあたりで，訂正後の認識率が逆転し，CN の方が改善されているのがわかる．認識率 80%になると，N-best では誤りを訂正できなくなるが，CN ではまだ誤りを訂正できている．

また，Fig. 3 は，Table 2 に基づいて，訂正された単語数を表している．ベースとなる単語認識率が高くなるにつれて，N-best と CN ともに訂正された単語数が減っていくが，認識率が 60%になった段階で，CN の訂正できる単語数が N-best を上回る．その後はベース認識率が 80%に到達するまで，CN の訂正できる単語数が上回っている．認識精度が低いうちは N-best の方が訂正能力が高かったが，現在の話し言葉における音声認識精度は前述の通り 80%前後なので，実験結果から CN の訂正能力が優れていることがわかる．

N-best が優れている範囲がある原因として，CN における CRF による誤り検出精度の低さが考えられる．Table 3 に，N-best について誤り検出する際の 1-best，CN について誤り検出する際の CN 最尤候補列，それぞれの誤り検出精度を示す．1-best の文における誤り検出精度が F 値で 0.686 であるのに対して，ヌル遷移を含む文における誤り検出精度は 0.578 となっている．提案した誤り訂正手法は，この誤り検出が基本となっているため，認識精度が低いうちは N-best に負けてしまう．CN には多くのヌル遷移が現れるため，表層単語 N-gram を適切に求められないことが，誤り検出精度を下げている原因だと考えられる．そのため，CN の第一候補を学習する際に，ヌル遷移はスキップし，訂正段階ではヌル遷移が含まれる CS に

Table 2　単語認識率ごとの訂正能力

	単語認識率 [%]					
base	30.00	40.00	50.00	60.00	70.00	80.00
N-best	34.99	43.91	53.00	62.06	70.94	79.96
Confusion Network	33.02	42.73	52.40	62.12	71.74	81.34

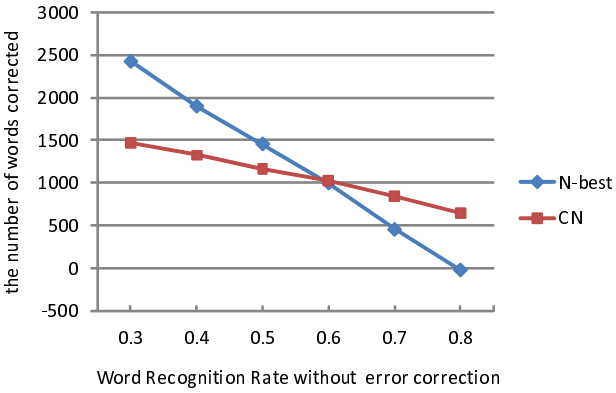

Fig. 3　訂正できた単語数

ついて，ヌル遷移を除いて誤り検出を行い，正解が見つからなければヌル遷移を選択する，というようなアルゴリズムを組み込めば改善するのではないかと考えられる．

5　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．様々な素性を自由に取り入れ，さらに単語ごとに訂正が行えるため，柔軟な誤り訂正が可能になった．この提案手法を用いて，日本語話し言葉コーパスによる評価実験で，単語認識率の改善が確認された．

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．そのために，4 章で述べたアルゴリズムの他に，CRF で学習する際の素性として品詞情報，連体形の後には体言しか現れないといった，活用形-品詞の連鎖情報や，助動詞の後に動詞が現れることはめったにないといった，品詞の bigram，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，高精度なパラメータ推定を行う [10] ことや，CRF の改良手法 [11] による学習を取り入れることも考えたい．

Table 3　誤り検出精度

	F 値
1-best	0.686
CN の最尤候補列	0.578

参考文献

[1] 中川 聖一, “音声ディクテーションから音声ドキュメント処理へ”, 音講論 (秋), pp. 1-4, 2007.

[2] B. Roark, *et al.* “Discriminative language modeling with conditional random fields and the perceptron algorithm,” ACL, pp. 47-54, 2004.

[3] 大庭, 他, “単語誤り率を考慮した誤り訂正モデル学習とその効果に関する分析”, 音講論 (春), pp. 127-128, 2008.

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

[5] 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.

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

[7] 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.

[8] 緒方, 後藤, “音声訂正：選択操作による効率的な誤り訂正が可能な音声入力インタフェース”, 情報処理学会論文誌, Vol. 48, No. 1, pp. 375-385, 2007.

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

[10] C. White, *et al.* “MAXIMUM ENTROPY CONFIDENCE ESTIMATION FOR SPEECH RECOGNITION,” ICASSP, pp. 809-812, 2007.

[11] P. Jian, *et al.* “Conditional Neural Fields,” NIPS22, pp. 1419-1427, 2009.