未知語モデルを用いたCRFに基づく音声認識誤り訂正*

☆中谷良平 (神戸大), 岩橋直人 (NICT), 中野幹生 (HRI-JP), 滝口哲也, 有木康雄 (神戸大)

1　はじめに

現在までに，音声認識技術は目覚しい発展を遂げてきた．しかし，誤認識を完全に防ぐことは依然として不可能である．この問題を解決するために，識別的言語モデルを用いて，大語彙連続音声認識によって出力されたN-best候補をリランキングする，音声認識誤り訂正技術が提案されている [1][2]．これらの手法は既知語の認識誤りを訂正することは可能だが，辞書に存在しない単語はN-bestリストに出現しないため，未知語の認識誤りを訂正することはできない．

そこで本稿では，hybrid word/syllable recognitionを実装することで，音声認識器に未知語認識機能を追加し，その後に音声認識誤り訂正を行う手法を提案する．この手法を用いることで，未知語を既知語として認識する誤りだけでなく，既知語を未知語として認識する誤りについても訂正することを目的としている．音声認識誤り訂正には以前提案していた手法 [3] を使う．これはConfusion Network [4] を競合仮説として扱い，Conditional Random Fields (CRF) [5] を用いて単語ごとに誤り訂正を行う手法である．

以降2章では，提案手法の流れについて述べる．3章では未知語認識手法について，4章では音声認識誤り訂正手法についてそれぞれ述べる．5章で評価実験とその結果を示し，6章でまとめについて述べる．

2　提案手法の流れ

Fig. 1は提案手法の流れを示している．点線で囲まれた学習プロセスでは，まず，hybrid word/syllable recognitionを行い，認識結果をConfusion Networkとして出力する．そして対応する書き起こしデータを用いてConfusion Network内の全単語に正誤ラベリングを行い，単語N-gram，Confusion Network上の存在確率などを素性として，CRFによって誤り検出モデルを学習する．Fig. 1下部の誤り訂正プロセスでは，hybrid word/syllable recognitionを行いConfusion Networkを生成する．そして，誤り検出モデルを用いて，Confusion Network上で単語ごとに正解を探す．

3　未知語認識

本稿では，クラスN-gramを用いて未知語認識を行う．n 個の単語からなる単語列 $\{w_1, \cdots, w_n\}$ が与えられたとき，一般的な単語N-gramは式 (1) のように定義される．

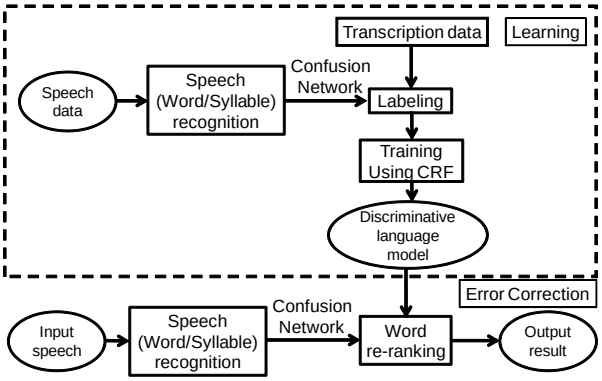

Fig. 1　提案手法の流れ

$$P(w_n|w_1, \cdots, w_{n-1}) = P(w_n|w_{n-N+1}, \cdots, w_{n-1}) \quad (1)$$

一方, クラスN-gramは式 (2) のように定義される．

$$\begin{aligned} &P(w_n|w_1, \cdots, w_{n-1}) \\ &\quad = \quad P(c_n|c_{n-N+1}, \cdots, c_{n-1})P(w_n|c_n) \end{aligned} \quad (2)$$

クラス $\{c_1, \cdots, c_n\}$ は，単語列 $\{w_1, \cdots, w_n\}$ のそれぞれの単語が属する単語クラスである．クラスN-gram確率 $P(c_n|c_{n-N+1}, \cdots, c_{n-1})$ はテキストデータから学習される．Fig. 2はクラスN-gramを用いた未知語認識手法の概要を示している．未知語クラス c_{OOV} に全ての未知語が属すると定義する．c_{OOV} には全ての音節が属していて，音節の連鎖は c_{OOV} の連鎖とすることで擬似的にFig. 2の構造を表現する．つまり，未知語は c_{OOV} の連鎖として表現され，c_{OOV} を含むクラスN-gramは未知語が出現しやすい場所を示している．Fig. 2では c_{OOV} の前後に c_i と c_j を含むクラスN-gramになっているので，c_i は未知語の前に現れやすいクラス，c_j は未知語の後ろに現れやすいクラスとなる．

c_{OOV} には音節しか属していないため，$P(w_n|c_{OOV})$ は c_{OOV} からそれぞれの音節が発生する確率となる．w_n として全ての音節を登録し，$P(w_n|c_{OOV})$ はどの音節に対しても等確率とする．この $P(w_n|c_{OOV})$ をパラメータとして変化させながら実験を行う．c_{OOV} を除いたすべてのクラスは1クラス1単語しか属さないように定義し，既知単語と未知語クラスのN-gramを

*Correcting Speech Recognition Errors Based on CRF Using Out-Of-Vocabulary Word Modeling, by Ryohei Nakatani (Kobe University), Naoto Iwahashi (NICT), Mikio Nakano (HRI-JP), Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

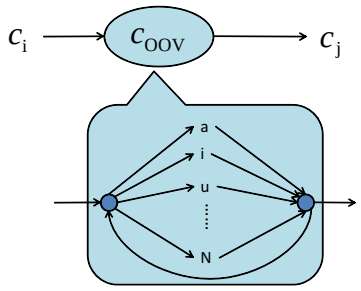

Fig. 2　クラス N-gram による未知語モデリング

Table 1　Hybrid word/syllable recognition の例

Input speech: “私たちは Perl スクリプトを使う”
The result of hybrid word/syllable recognition: “私 たち は Syl_pa: Syl_ru スクリプト を 使う”

テキストから学習することで，hybrid word/syllable recognition を実現している．

Table 1 は “Perl” を未知語としたときの hybrid word/syllable recognition の例である．“私たちは Perl スクリプトを使う” という発話に対して hybrid word/syllable recognition を行うと，“私 たち は Syl_pa: Syl_ru スクリプト を 使う” という出力が得られた．ここで，“Syl_pa:” と “Syl_ru” はそれぞれ “pa:”，“ru” という発音を表す音節であり，コロン “:” は長音記号を表している．このように，既知語部分を単語列で，未知語部分を音節列で出力するのが hybrid word/syllable recognition の目的である．

4　音声認識誤り訂正

4.1　Conditional Random Fields

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) [5] でモデル化する．CRF を用いた誤り検出モデルは，音声認識結果とそれに対応する書き起こしデータを用いて学習され，入力文書中の不自然な単語を検出することができる．

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．

$$P(y \mid x) = \frac{1}{Z(x)} \exp\Big(\sum_a \lambda_a f_a(y, x)\Big) \tag{3}$$

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．

$$Z(x) = \sum_y \exp\Big(\sum_a \lambda_a f_a(y, x)\Big) \tag{4}$$

パラメータ λ_a は，学習データ $(x_i, y_i)(1 \le i \le N)$ が

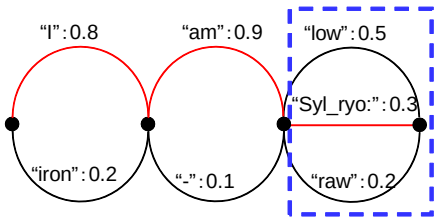

Fig. 3　Confusion Network の例

与えられたとき，条件付確率分布 (3) の対数尤度，

$$\mathcal{L} = \sum_{i=1}^{N} \log P(y_i \mid x_i) \tag{5}$$

を最大にするように学習される．学習は，準ニュートン法である L-BFGS 法 [6] によって行われる．

4.2　Confusion Network

提案しているシステムでは，CRF によって音声認識誤りを検出し，他の競合仮説と置き換えることで誤り訂正を行う．本稿では，競合仮説の表現として Confusion Network を用いる [4]．

Fig. 3 は “I am Ryo” という発話を認識した際の Confusion Network の例である．点線で囲まれた部分は信頼度が付与された競合単語候補として表現されていて，Confusion Set と呼ばれる．Fig. 3 中には 3 つの Confusion Set が描かれている．信頼度の最も高い候補を選択していくと最尤候補となり，Fig. 3 の例では “I am low” となる．“-” で表された遷移はヌル遷移と呼ばれ，候補単語が存在しないことを意味している．

例えば，Fig. 3 の 3 番目の Confusion Set には，“low”，“Syl_ryo:”，“raw” の 3 つの競合仮説が存在する．ここで，“Syl_ryo:” は “ryo:” という音節を意味する．最も尤度の高い単語列は “I am low” となるが，CRF によって “low” という単語を誤りだと識別することが出来れば，第 2 候補である “Syl_ryo:” と置き換えられる．

4.3　誤り訂正アルゴリズム

前節で述べたように，本稿では CRF を用いて誤り訂正を行う．普通，CRF による誤り傾向の学習には音声認識結果の 1-best 単語列を用いるが，本稿で用いる Confusion Network には特有のヌル遷移が多数存在するため，Confusion Network の第一候補単語列（最尤候補），第二候補単語列，第三候補単語列に正誤ラベリングしたものを，CRF によって学習する．ここで，第三候補がない Confusion Set については，第二候補で補い，第二候補がない Confusion Set については，第一候補で補っている．また，学習に用いる素性は，次章で述べる．誤り検出モデルの学習後，以下のアルゴリズムに従って誤り訂正を行う．

Table 2　データ数

	Training	Test
Number of lectures	106	4
Number of speeches	50,780	1,618
Number of words	513,281	17,594

1. 評価データを音声認識後，Confusion Network を出力する．
2. Confusion Network の第一候補列のみを抜き出し，CRF による誤り検出を行って，正誤ラベルを付与する．
3. 入力時系列順に Confusion Set を見ていく．正解と判定された語には何も操作を行わずに次の Confusion Set へ進む．誤りと判定された語は，対応する Confusion Set から次の候補を選び出し，置き換えてもう一度 CRF による誤り検出を行う．
4. Confusion Set の中に正解と識別された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 Confusion Set 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,「入力時系列順に」と述べたのは，CRF によって学習する際の素性として単語 N-gram 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 単語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．

5　評価実験

5.1　実験条件

本研究ではベースとなる音声認識システムに，大語彙連続音声認識エンジン Julius-4.1.4 [7] を用いる．

音響モデルは，CSJ の学会講演のうち，953 講演（男性 787 講演+女性 166 講演），計 228 時間分の講演音声から作成した HMM を用いた．言語モデルは，CSJ の書き起こし文書のうち，2,596 講演の書き起こし文書から学習した N-gram を用いた．

また，学習と評価に用いたデータ数を Table 2 に示す．学習には 106 講演分，評価には 4 講演分の音声データをそれぞれ用いた．コーパスは CSJ を用いている．学習には，Julius が出力した Confusion Network を用いた．誤り傾向を学習するための素性は，表層単語 unigram，bigram，trigram，そして Confusion Netwrok 上の存在確率を用いた．

実験は，最適なパラメータ値を調査し，hybrid word/syllable recognition の精度をより良くするために，3 章で述べた音節の発生確率を制御するパラメータ $P(w_n|c_{OOV})$ を，−0.01 から −5.0 まで変化させながら行う．また，未知語は単語辞書に含まれる名詞からランダムに選ばれており，選ばれた単語を辞書から削除した．テストデータに含まれる未知語の割合は 2 %である．

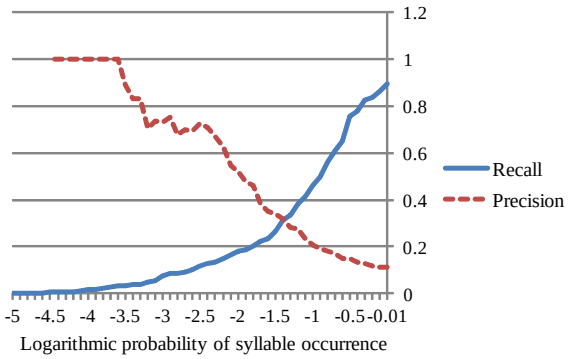

Fig. 4　Recall-Precision 曲線

5.2　実験結果

Fig. 4 に未知語検出の Recall-Precision 曲線を示す．横軸の値は 3 章で述べた未知語音節の発生確率である．このパラメータは未知語の発生しやすさを制御している．つまり，この値が大きいほど認識器が未知語を出力しやすくなる．Recall と Precision は次のように定義している．

$$Recall = \frac{The\ number\ of\ truly\ recognized\ OOV\ words}{The\ number\ of\ true\ OOV\ words\ in\ the\ reference\ file} \quad (6)$$

$$Precision = \frac{The\ number\ of\ truly\ recognized\ OOV\ words}{The\ number\ of\ words\ recognized\ as\ OOV\ words} \quad (7)$$

Recall 曲線と Precision 曲線は，値の変動が急激になってしまっているが，パラメータの値が増加するとともに Recall が増大し Precision が減少しているため，意図した通りの挙動となっている．用いた hybrid word/syllable recognition が単純すぎたため，あまり高いパフォーマンスを得ることができていない。

Fig. 5 は，パラメータを変化させた場合の音声認識結果を示している．評価指標としては，単語誤り率（Word Error Rate：WER）を用いている．未知語については，音節系列が異なっていても，場所さえ合っていれば正解とした．“Base” は従来通りの音声認識結果で，2 %の未知語を全て誤認識してしまう．“OOV modeling” は同じく 2 %の未知語を含んでいて，3 章で述べた hybrid word/syllable recognition を用いて得られた認識結果である．また，“Oracle” は音声データに現れる全ての単語を既知語にした場合の音声認識結果である．横軸のパラメータは未知語の発生しやすさを制御するため，その値が小さくなればなるほど未知語を検出しにくくなり，“OOV modeling” は徐々に “Base” に近づいていく．パラメータが −4.9 以下のとき，未知語を認識しなくなり，“Base” と “OOV modeling” は一致する．また，パラメータが −0.01 に近づくほど未知語が過剰に発生するようになり，未知語の挿入誤りが増えるために WER は高くなって

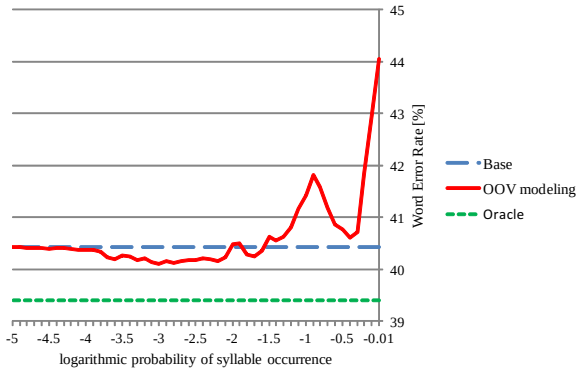

Fig. 5　単語誤り率

Table 3　誤りタイプごとの評価

	SUB	DEL	INS	COR	WER
Base	4,213	813	2,086	10,045	40.42
OOV modeling	4,174	796	2,112	10,132	40.25
Base corrected	4,114	927	1,797	10,030	38.86
Proposed method	4,069	905	1,776	10,128	38.37

いく．パラメータの値が −1.7 付近のときと −2.1 から −4.9 の間のとき，“OOV modeling” の WER が “Base” より良くなっている．

Table 3 は，単語誤り率と誤りタイプごとの誤り数となっている．それぞれ，“SUB” は置換誤り，“DEL” は削除誤り，“INS” は挿入誤り，“COR” は正解単語の数である．Fig. 5 において，パラメータの値が −1.7 のときの単語誤り率を，Table 3 の “OOV modeling” として固定した．“Base corrected” は 4 章で述べた誤り訂正手法に基づいて “Base” の誤り訂正を行った結果である．同様に，“Proposed method” は “OOV modeling” の誤りを訂正した場合の結果となっている．提案手法の置換誤りと挿入誤りの数は最も小さくなっていて，結果として，単語誤り率も最も小さくなっている．“Base” と比較すると，40.42 %から 38.37 %まで低下し，トータルで 2.05 ポイント改善した．“Base corrected” と比較しても 0.49 ポイントの改善となった．

また，Table 4 は誤り訂正前と訂正後の Recall，Precision の変化を表している．Recall は変化していないものの，Precision が大きくなっている．このことから，訂正後は訂正前と比べて正しく認識できた未知語はそのままに，既知語を未知語に誤認識してしまうケースが減ったことがわかる．その理由としては，未知語検出精度が低いため，既知語を未知語として誤認識してしまうケースが多く，学習段階で未知語自体が誤り傾向を示す語として学習されてしまったことが考えられる．

Table 4　訂正前と訂正後の Recall，Precision の変化

	Before	After
Recall	0.22	0.22
Precision	0.38	0.40

6　まとめ

本稿では，未知語モデルを利用して hybrid word/syllable recognition を行った後に誤り訂正を行うことで，今まで訂正不可能だった音声認識誤りの訂正を可能にした．単語誤り率は 40.42 %から 38.37 %まで改善した．これは，ベースラインを誤り訂正した場合と比較しても 0.49 ポイントの改善である．

今回用いた未知語認識手法はシンプルで実装が簡単だが，効果はあまり大きくなかったため，今後の課題として，Rastrow らが提案している未知語検出手法 [8] など，もっと効果的な未知語認識手法を導入することが挙げられる．また，CRF で学習する際の素性として，文脈を考慮した N-gram よりも広範囲な長距離言語情報などを用いることも考えたい．

参考文献

[1] B. Roark, M. Saraclar, M. Collins, and M. Johnson, “Discriminative language modeling with conditional random fields and the perceptron algorithm,” in *Proc. ACL*, pp. 47–54, 2004.

[2] T. Oba, T. Hori, and A. Nakamura, “A study of efficient discriminative word sequences for reranking of recognition results based on n-gram counts,” in *Proc. Interspeech2007*, pp. 1753–1756, 2007.

[3] 中谷良平, 滝口哲也, 有木康雄, “文脈特徴を用いた CRF による音声認識誤り訂正”, 日本音響学会講演論文集（秋）, pp. 189–190, 2011.

[4] L. Mangu, E. Brillx, and A. Stolcke, “Finding consensus in speech recognition: word error minimization and other applications of confusion networks,” in *Computer Speech and Language*, pp. 373–400, 2000.

[5] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proc. ICML*, pp. 282–289, 2001.

[6] J. Nocedal, “Updating quasi-newton matrices with limited storage,” in *Mathematics of Computation*, pp. 773–782, 1980.

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

[8] A. Rastrow, A. Sethy, and B. Ramabhadran, “A new method for oov detection using hybrid word/fragment system,” in *ICASSP2009*, pp. 3953–3956, 2009.