Syntax 情報と Context 情報を用いた音声認識誤りの 2 段階訂正*

☆中谷良平, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

本稿では，単語ごとに付与した長距離文脈スコアを素性とし，Confusion Network 上で音声認識自動誤り訂正を行う手法を提案する．従来，単語ごとに付与された長距離文脈情報を素性として音声認識誤り訂正を行う手法は提案されているが [1]，単語ごとにそれを付与する場合，周辺の認識精度に大きく依存してしまうという問題があった．そのため，認識誤りを多く含む認識結果に対して長距離文脈情報を付与することは，あまり好ましくない．したがって本研究では，長距離文脈情報を誤り訂正の素性として用いるために，始めに N-gram 情報を用いた誤り訂正を行い，誤認識を軽減する．その後，長距離文脈スコアを付与し，2 段階目の訂正を行うことで，音声認識精度を向上させる手法を提案する．実験により，提案する 2 段階訂正を行うことで，より効果的に長距離文脈情報を誤り訂正の素性として利用できることを確認した．

2 提案手法の流れ

Fig. 1 は提案手法の流れを示している．左上の点線で囲まれた Learning N-gram model プロセスは，N-gram や認識信頼度などの構文情報を用いた誤り検出モデルの学習プロセスである．まず，通常の音声認識を行い，認識結果を Confusion Network [2] 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 内のすべての単語に正誤ラベリングを行い，bigram，trigram，Confusion Network 上の存在確率などを素性として，Conditional Random Fields (CRF) [3] により誤り検出モデルを学習する．

また，中段の点線で囲まれた Learning context model プロセスは，Context，つまり長距離文脈情報を用いた誤り検出モデルの学習プロセスである．先ほどとは異なる発話データについて，音声認識後，既に学習済みの N-gram 情報を用いた誤り検出モデルを用いて誤り訂正を行う．そうして可能な限り認識誤りを削減した後，LSA を用いて単語ごとに文脈スコアを付与する．その後は同様に正誤ラベリングを行い，先ほどの素性に加え，Latent Semantic Analysis (LSA) [4] による文脈スコアを素性として，CRF によって誤り検出モデルを学習する．

Fig. 1 下部の Test プロセスは評価実験の処理である．始めに，音声データに対して音声認識を行い Confusion Network を生成する．そして，1 ステップ目として N-gram 情報から学習した誤り検出モデルを用いて，Confusion Network 上で単語ごとに誤り訂正を行う．その後，文脈スコアを計算し，Context 情報から学習した誤り検出モデルを用いて，2 ステップ目の誤り訂正を行う．

3 長距離文脈情報

3.1 長距離文脈スコア

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どれだけ自然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. 2 のように,「音声」「会話」「話者」「対話」などの単語が含まれる話題の中に,「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長距離文脈スコアとして算出し，誤り検出に用いる．しかし，長距離文脈スコアは，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 の長距離文脈スコア，$SC(w)$ は次のように計算する．w の周辺に現れる内容語を，Fig. 2 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)．$c(w)$ 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．

$$SC(w_i) = sim(w_i, c(w)) \tag{1}$$

単語間類似度 $sim(w_i, c(w))$ の算出には LSA を用いた．

3.2 Latent Semantic Analysis

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場合でも，単語間の類似度を求めることができる手法である [4].

学習手順としてはまず，N 個の文書から単語文書行列 W を生成する．本研究では W の要素 w_{ij} として tf-idf を用い，以下のように求める．

$$w_{ij} = tf_{ij} \cdot idf_i \tag{2}$$

*Two-step Correction of the Speech Recognition Result based on Syntax and Context Information, by Ryohei Nakatani, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

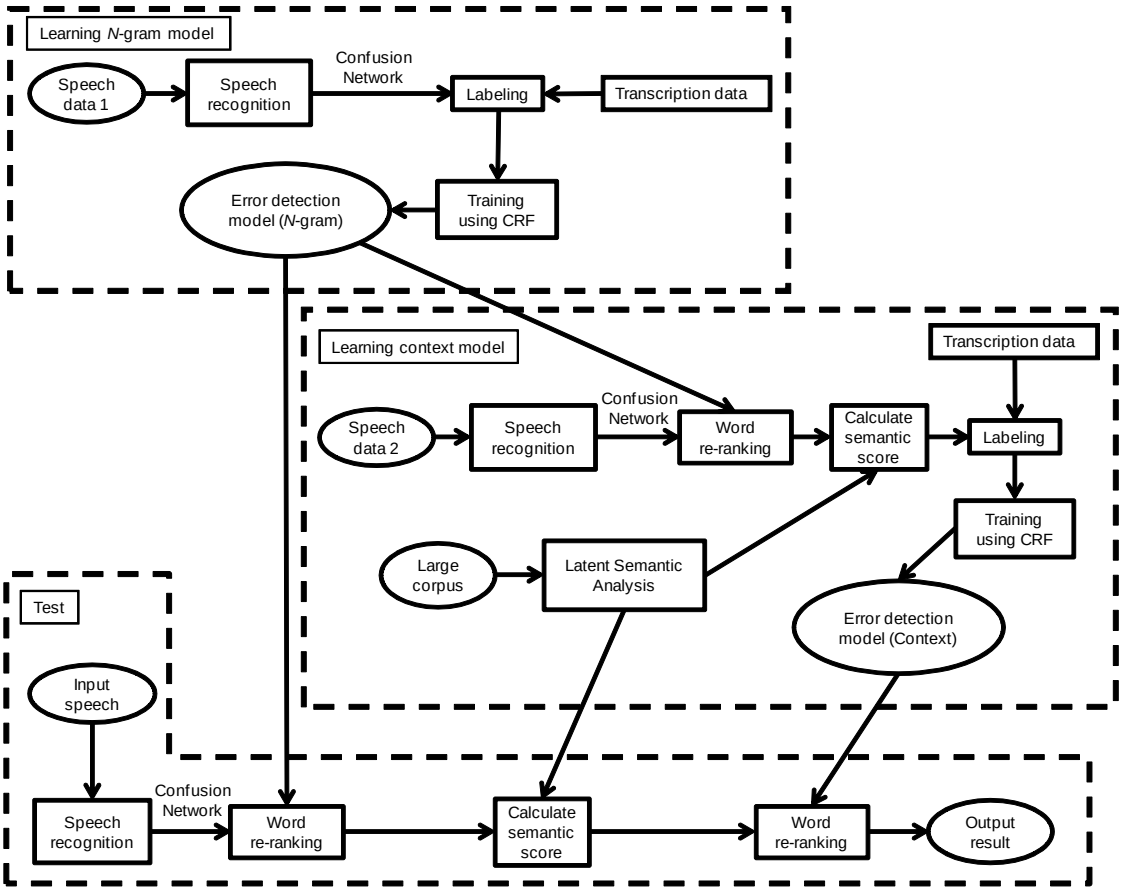

Fig. 1　Flow of proposed method

$$tf_{ij} = \frac{n_{ij}}{|\ c_j\ |} \tag{3}$$

$$idf_i = \log \frac{N}{df_i} \tag{4}$$

tf-idf は単語の出現頻度を表す tf と，逆出現頻度を表す idf の 2 つの指標で計算される。n_{ij} は文書 c_j における単語 r_i の出現回数，$|\ c_j\ |$ は文書 c_j に含まれる単語の総数，df_i は単語 r_i が出現する文書の総数である．idf_i は単語 r_i の単語重みと考えることができ，多くの文書で出現する単語では小さく，特定の文書でしか出現しない単語では大きくなるという特徴がある．

ここで，語彙数を M とすると，行列 W は $M \times N$ のスパースな行列となる．そのため，この単語文書行列 W を特異値分解し，特異値の大きなものから $R(< rank(W))$ だけ用いることで次のような近似を行う．

$$W \approx \hat{W} = USV^T \tag{5}$$

特異値分解により各行列は Fig. 3 のような形になっている．$U(M \times R)$ は各単語 r_i に対応する行ベクトル $u_i(1 \leq i \leq M)$ から成る単語行列，$S(R \times R)$ は特異値の対角行列，$V(N \times R)$ は各文書 c_j に対応する行ベクトル $v_j(1 \leq j \leq N)$ から成る文書行列である．この次元圧縮により，関連の強い単語は同一次元に縮約され，直接経験したことのない単語の共起関係についても，類似度を得ることができる．

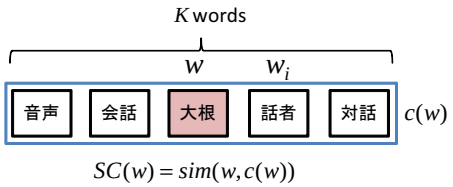

Fig. 2　Calculation of semantic score

LSA では単語 r_i と文書 c_j の類似度 $sim(r_i, c_j)$ を，以下の式により求める．

$$sim(r_i, c_j) = \frac{u_i S v_j^T}{\| u_i S^{\frac{1}{2}} \| \| v_j S^{\frac{1}{2}} \|} \tag{6}$$

$sim(r_i, c_j)$ は 1 に近いほど類似度が高く，-1 に近いほど類似度が低いことを示す．

4　CRF を用いた音声認識誤り訂正

本稿では誤り検出モデルを，音声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 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

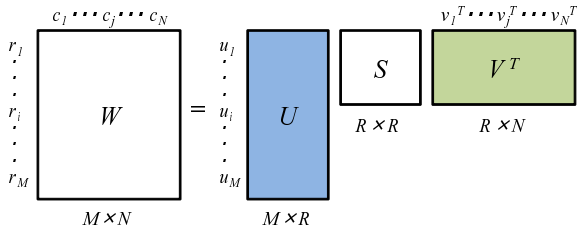

Fig. 3　Latent Semantic Analysis

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 は系列データを学習データとするため，本稿では，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．

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 の中に正解単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，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 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 [5] を用いる．

音響モデルは，日本語話し言葉コーパス (CSJ) の学会講演のうち，953 講演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．

Table 1　The number of data

	N-gram Training	Context Training	Test
Number of lectures	150	150	301
Number of words	259,901	311,374	113,289

Table 2　Features used for error tendency learning

	N-gram model	Context model
Unigram	○	○
Bigram	○	○
Trigram	○	○
Confidence of Confusion Network	○	-
Long-distance context score	-	○

また，本稿では Fig. 1 が示すように，LSA の学習データ，N-gram 情報を用いた誤り検出モデルの学習データ，そのモデルを用いて誤り訂正後，長距離文脈スコアを付与し，Context 情報を用いた誤り検出モデルを学習するためのデータ，評価データの 4 つのデータセットが必要になる．各データセットについて以下に示す．

LSA の学習には，CSJ の書き起こし文書，2,672 講演分のデータを用いた．内容語として名詞，動詞，形容詞のみを扱い，語彙数は 48,371 であった．内容語が 30 語程度出現するごとに区切った区間を文書の単位とし，文書数は 76,767 となった．特異値分解では 100 次元に圧縮した．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．

N-gram 誤り検出モデル，Context 誤り検出モデルのそれぞれの学習と，評価に用いたデータ数を Table 1 に示す．N-gram 誤り検出モデルの学習には 150 講演分の音声データ，Context 誤り検出モデルの学習にはそれと異なる 150 講演分の音声データ，評価には学習データを含まない 301 講演分の音声データをそれぞれ用いた．

次に，誤り検出モデルを学習するための素性を Table 2 に示す．どちらのモデルも，単語 unigram，bigram，trigram を素性としている部分は共通している．N-gram model と Context model の違いは，N-gram と同時に学習する素性として，Confusion Network 上の信頼度か，長距離文脈スコアのどちらを選択するかである．また，Table 1 が示すように，異なるデータセットを用いてそれぞれのモデルを学習している．

Table 3　Evaluation with each error type

	N-gram model	Context model	SUB	DEL	INS	COR	WER [%]
Recognition result	×	×	28,446	5,453	14,751	63,871	42.94
N-gram model	○	×	21,322	7,227	8,971	69,221	33.12
Context model (Baseline)	×	○	21,267	7,072	9,070	69,431	33.02
2 step correction (N-gram)	○ ○	×	19,132	9,193	6,374	69,445	30.63
Proposed method	○	○	18,144	10,052	5,203	69,574	29.48

5.2　実験結果

Table 3 は，単語誤り率と誤りタイプごとの誤り数を表している．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．“Recognition Result” は，Test データセットを音声認識した際の結果である．“N-gram model” と “Context model (Baseline)” は共に，Table 1 の学習データを全て (300 講演) 用いて誤り検出モデルを学習している．用いた素性は以下の通りである．“N-gram model” は，単語 N-gram と認識信頼度を用いて，“Recognition Result” を誤り訂正した結果である．同様に，“Context model (Baseline)” は N-gram 特徴に加えて，長距離文脈スコアを素性として学習したモデルを用いて誤り訂正を 1 段階で行った結果である．ただし，このモデルについては，本稿の “周辺単語に認識誤りが少ないほど，長距離文脈情報が効果的に利用できる” という提案と比較するために，従来同様，学習データにリランキングを行わず，認識結果にそのまま長距離言語スコアを付与したものから学習している．

“Proposed method” は，提案手法である Fig. 1 に従って実験を行った結果である．Table 3 には，N-gram model と Context model のうち，どれを使用しているかが分かるよう，○ (使用)，× (未使用) を表示している．また，“2 step correction (N-gram)” は，“Proposed method” と同じ 2 段階訂正だが，2 つの誤り検出モデルをいずれも Table 2 の N-gram model の素性を用いて学習している．つまり，異なるデータから，同じ素性で 2 つの誤り検出モデルを学習した．これは，1 段階訂正と 2 段階訂正を比較するためのものである．すなわち，1 段階の従来手法で，長距離文脈情報を用いた場合の改善 (“N-gram model” と “Context model” の WER の差) と，提案した 2 段階訂正で長距離文脈情報を用いた場合の改善 (“2 step correction” と “Proposed method” の WER の差) を比較するために用意した．

Table 3 が示すように，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．“Baseline” と比較すると，33.02 %から 29.48 %まで低下し，トータルで 3.54 ポイント改善した．誤り削減率では，10.72 %を達成した．また，“N-gram model” と従来の “Context model” を比較すると，認識誤りに影響され，長距離文脈スコアを効果的に用いることができていなかったために，WER が 0.1 ポイントの改善，0.3 %の誤り削減しか達成できていなかった．しかし，“2 step correction” と “Proposed method” を比較すると，WER が 1.15 ポイント改善し，3.75 %の誤り削減を達成した．この結果から，認識誤りを削減してから長距離文脈スコアを計算することで，従来手法よりも効果的に長距離文脈情報を利用できることが示せた．

6　おわりに

本稿では，長距離文脈上情報を音声認識誤り訂正における素性の一つとして用いるために，1 段階目で N-gram 情報を用いて誤り訂正を行うことで認識誤りを可能な限り削減し，その後，長距離文脈スコアを付与して 2 段階目の誤り訂正を行う手法を提案した．認識誤りを多く含む通常の認識結果に長距離文脈情報を付与して訂正した従来手法と，本提案手法とを比較すると，10.72 %の誤り削減を達成した．また，従来は長距離文脈情報を追加しても，0.3 %の誤り削減しか達成できなかったが，提案した 2 段階訂正では，長距離文脈情報を追加した場合，3.75 %の誤り削減を達成した．

今後の課題として，CRF を改善した手法である Conditional Neural Fields [6] を利用することや．LSA 以外の単語間類似度の計算方法も考えていきたい．

参考文献

[1] 中谷良平, 他, “文脈特徴を用いた CRF による音声認識誤り訂正,” 音講論（秋）, pp. 189–190, 2011.

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

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

[4] T. Landauer, *et al.* “Introduction to Latent Semantic Analysis,” in *Discourse Processing*, pp. 259–284, 1998.

[5] Julius development team, “大語彙連続音声認識エンジン Julius,” http://julius.sourceforge.jp/.

[6] J. Xu, *et al.* “Conditional neural fields,” in *Proc. NIPS2009*, pp. 1419–1427, 2009.