一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2012-96(2012-12)

シンタックスとセマンティックスに基づく音声認識結果の2段階訂正

中谷　良平†　滝口　哲也††　有木　康雄††

† 神戸大学システム情報学研究科　〒657–8501 兵庫県神戸市灘区六甲台町 1–1
†† 神戸大学自然科学系先端融合研究環　〒657–8501 兵庫県神戸市灘区六甲台町 1–1
E-mail: †nakatani@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし　本稿では，単語ごとに長距離文脈スコアを付与することで素性とし，Confusion Network 上での音声認識自動誤り訂正手法を提案する．従来，単語ごとの長距離文脈情報を素性に音声認識誤り訂正を行う手法は提案されているが，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る．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のは，あまり好ましくない．したがって本稿では，文脈情報を誤り訂正の素性として用いるために，まずはシンタックス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，より音声認識精度を向上させることを目的とする．
キーワード　confusion network, conditional random fields, 音声認識誤り訂正, 文脈情報

Two-step Correction of the Speech Recognition Result based on Syntax and Semantics

Ryohei NAKATANI†, Tetsuya TAKIGUCHI††, and Yasuo ARIKI††

† Graduate School of System Informatics, Kobe University　1–1 Rokkodai, nada, Kobe 657–8501, Japan
†† Organization of Advanced Science and Technology, Kobe University　1–1 Rokkodai, nada, Kobe 657–8501, Japan
E-mail: †nakatani@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract　This paper presents the new method correcting speech recognition errors base on long-distance context. As in the past, the method which corrects recognition errors using long-distance context information given every word has been already proposed However, this method has the problem that a context score every word depends on peripheral recognition errors considerably. So, it is not desirable that long-distance context information is given the recognition result containing a lot of recognition errors. Therefore, in this paper, recognition errors are reduced by error correction adopting features of syntax to use context information as one of the feature. And then after correcting results are given long-distance context score, residual recognition errors are corrected by using that score as the feature.
Key words　confusion network, conditional random fields, word-error correction, semantics

1. まえがき

現在までに，音声認識技術は目覚しい発展を遂げてきた．アナウンサーが書き言葉で書かれた原稿を読み上げるような場合では，単語正解精度において 95 %程度の性能で認識が可能である [1]．また，学会講演音声のような，話し言葉・自由発話音声であっても，85 %程度の性能が得られるようになってきた．これらの成果から，音声メディアのアーカイブ化において多くの研究がなされている．例えば，World Wide Web 上のポッドキャストを対象に音声メディアのアーカイブ化を行う PodCastle [2] や，MIT の講義映像/音声を対象とした MIT Lecture Browser [3] などがあげられる．これらのシステムでは，Word Error Rate (WER) を低くすることが求められる．言語モデルは音響モデルによって推測された候補に従って，最適な単語列を選択することができるが，現在の音声認識では音声認識誤りを避けることは難しい．

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 [4] [5] [6] [7]．これらは間違った単語を含む音声認識結果の単語列を負例，対

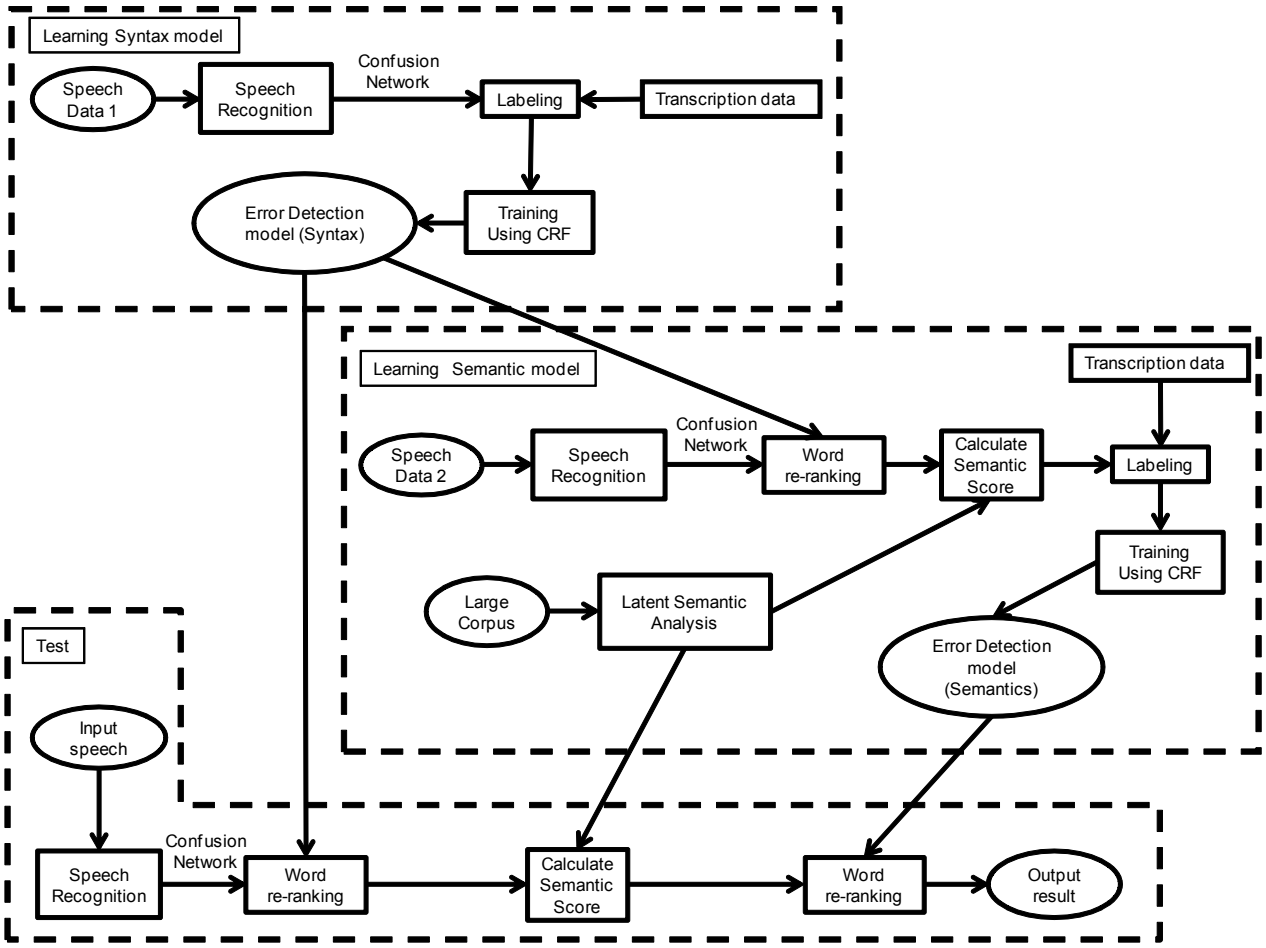

図 1　提案手法の流れ

Fig. 1　Flow of proposed method

応する書き起こしデータを正例として識別的に学習することで，N-best 単語列からより誤り特徴の少ない単語列を選び出す．また，我々は単語ごとに長距離文脈情報（ある単語が出現文中でどれだけ自然か）をスコア化し，誤り訂正の素性として用いてきた [8]．しかし，単語ごとに長距離文脈スコアを計算すると，周辺単語に非常に影響を受けるため，認識誤りが多いほど文脈情報の信頼性が低くなってしまうという問題がある．

そこで本稿では，文脈情報を誤り訂正でより効果的に用いるために，2 段階の訂正を行う手法を提案する．1 段階目の訂正では認識スコアや単語 N-gram などのシンタックスを用い，可能な限り認識誤りを訂正することで，正確に長距離文脈スコアを計算できるようにする．その後，長距離文脈スコアを付与し，新たな素性として 2 段階目の誤り訂正を行うことで，より効果的に文脈情報を扱えることを示す．

以降 2 章では，提案手法の流れについて述べる．3 章では長距離文脈情報ついて，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

図 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 Syntax 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 [9] 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) [10] によって誤り検出モデルを学習する．

また，右側の点線で囲まれた Learning Semantic model プロセスは，セマンティックス，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，既に学習済みの構文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，Latent Semantic Analysis (LSA) [11] 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，LSA 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．

図 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 ステップ目としてシンタックス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，セマンティックス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

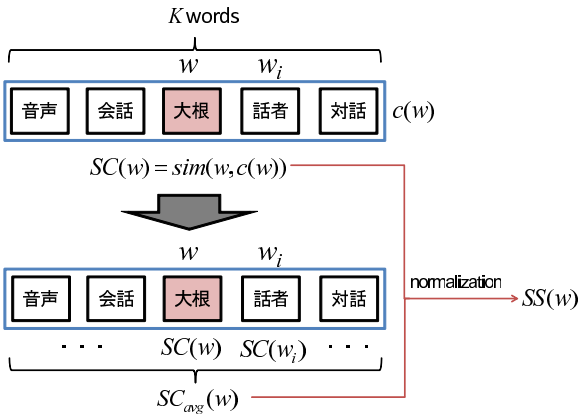

図 2　長距離文脈スコアの計算
Fig. 2　Calculation of semantic score

り訂正を行う．

3. 長距離文脈情報

3.1 長距離文脈スコアの計算

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．例えば図 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，$SS(w)$ は次のように計算する．

（1） w の周辺に現れる内容語を，図 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)．

（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．

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

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

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

（4） $SC(w)$ と $SC_{avg}(w)$ の差を長距離文脈スコア $SS(w)$ とする．

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

$SC(w)$ が大きいほど周辺に意味が近い単語が多いことになるが，強いトピックを持たない場合，$SC(w)$ は全体的に小さくなってしまう．そのため，$SC(w)$ に加え，SC_{avg} で正規化した $SS(w)$ の 2 つを長距離文脈スコアとして用いる．また，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 LSA を用いた．

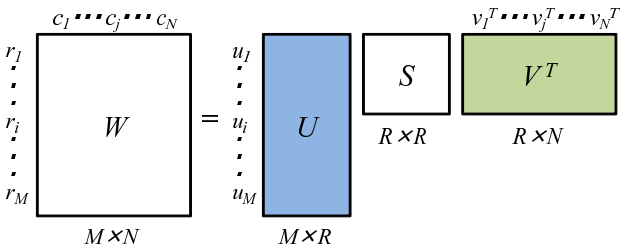

図 3　潜在的意味解析
Fig. 3　Latent Semantic Analysis

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 [11]．

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，以下の式により求める．

$$w_{ij} = tf_{ij} \cdot idf_i \tag{4}$$

$$tf_{ij} = \frac{n_{ij}}{|\, c_j \,|} \tag{5}$$

$$idf_i = \log \frac{N}{df_i} \tag{6}$$

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．

$$W \approx \hat{W} = USV^T \tag{7}$$

特異値分解により各行列は図 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 $u_i (1 \leqq i \leqq M)$ 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 $v_j (1 \leqq j \leqq N)$ 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．

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．

$$sim(r_i, c_j) = \frac{u_i S v_j^T}{\| u_i S^{\frac{1}{2}} \| \| v_j S^{\frac{1}{2}} \|} \tag{8}$$

$sim(r_i, c_j)$ 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．

ここで，入力文書を c_h としたとき，$v_h S^{\frac{1}{2}}$ を算出することを考える．LSA の学習に用いるコーパスが，潜在的意味空間を構成するのに十分な大きさのものであるならば，

$$c_h = USv_h^T \tag{9}$$

が成り立つと考えられる．この式を変形することで

$$v_h S^{\frac{1}{2}} = c_h^T U S^{-\frac{1}{2}} \tag{10}$$

が得られる．このようにして求まった $v_h S^{\frac{1}{2}}$ を利用することで，式 (8) の類似度の算出が可能となる．

4. 音声認識誤り訂正

4.1 Conditional Random Fields

Conditional Random Field (CRF) [10] は，主に自然言語処理やバイオインフォマティクスの分野で用いられているグラフ構造を持つ識別モデルである．文などの構造を持つデータ系列を扱い，モデル式は観測データ系列が与えられたときの出力ラベル系列の条件付確率分布という形をとる．ラベルが与えられた学習データ系列によってモデルを学習し，テストデータ系列を入力すると，モデルが推定するラベル系列が出力される．このとき，データ系列内の各データ一つ一つに最適と推定するラベルを割り当てるのではなく，系列全体として最適と推定するラベルを各データに割り当てる．これは，モデル学習時にデータ間の関係も学習し，ラベル推定時にデータ間の関係を考慮した上で，各データのラベルを推定することで実現する．

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，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．

$$P(y \mid x) = \frac{1}{Z(x)} \exp\Big(\sum_a \lambda_a f_a(y, x)\Big) \tag{11}$$

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．

$$Z(x) = \sum_y \exp\Big(\sum_a \lambda_a f_a(y, x)\Big) \tag{12}$$

パラメータ λ_a は，学習データ $(x_i, y_i)(1 \leqq i \leqq N)$ が与えられたとき，条件付確率分布 (11) の対数尤度，

$$\mathcal{L} = \sum_{i=1}^{N} \log P(y_i \mid x_i) \tag{13}$$

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 [12] によって行われる．

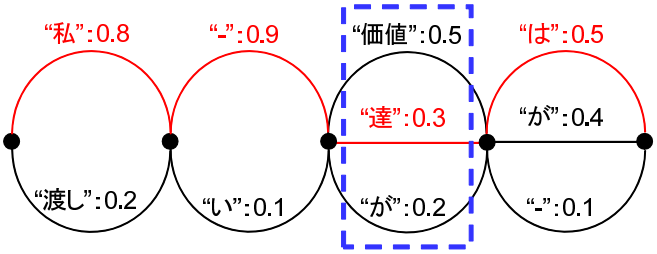

図 4　Confusion Network の例
Fig. 4　Example of Confusion Network

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．

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(y \mid x) \tag{14}$$

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，単語ごとの誤り訂正を行うために，競合仮説の表現として Confusion Network を用いる [9]．

Confusion Network 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．図 4 は “私達は” 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．図 4 中には 4 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，図の例では “私 価値 は” 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，図 4 の 3 番目の Confusion Set には，“価値”，“達”，“が” の 3 つの競合仮説が存在する．最も尤度の高い単語列は “私 価値 は” となるが，CRF によって “価値” 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 “達” 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，より多くの学習データを利用するために，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 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 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 [13] 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．音響分析条件と HMM の仕様は表 1 のようになっている．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．N-gram エントリーは表 2 のようになっている．

また，本稿では図 1 が示すように，LSA の学習データ，シンタックス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，セマンティックス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．ここで，本実験で用いるデータは全て日本語話し言葉コーパス（CSJ）のものである．

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．

シンタックス誤り検出モデル，セマンティックス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表 3 に示す．シンタックス誤り検出モデルの学習には 150 講演分の音声データ，セマンティックス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．Confusion Network は Julius によって出力している．

次に，誤り傾向を学習するための素性を表 4 に示す．どちらのモデルも，単語 unigram，bigram，trigram，Confusion Network 上の信頼度を素性としている部分は共通している．Syntax model と Semantic model の違いは，長距離文脈スコアを素性として加えたかどうかである．また，表 3 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．

表 1 音響分析条件と HMM の特徴
Table 1 Speech analysis conditions and specifications of HMM

Sampling frequency	16 kHz
Acoustic feature	MFCC (12 dim.) +Δ MFCC (12 dim.)
	+Δ power (total 25 dim.)
Window type	Hamming window
Frame length	25 ms
Frame shift length	10 ms
Acoustic model	Triphone (3,000 states)
The number of mixtures	16
State	5 states and 3 loops

表 2 N-gram エントリー
Table 2 The number of N-gram entries

Unigram	Bigram	Trigram
25,300	731,728	2,611,952

表 3 学習，評価データ数
Table 3 The number of data

	Syntax Training	Semantic Training	Test
Number of lectures	150	150	301
Number of words	259,901	311,374	113,289

表 4 誤り傾向の学習に用いる素性
Table 4 Features used for error tendency learning

	Syntax model	Semantic model
Unigram	○	○
Bigram	○	○
Trigram	○	○
Confidence of Confusion Network	○	○
Long-distance context score	-	○

5.2 実験結果

表 5 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．“Recognition Result” は，Test データセットを音声認識した際の結果となっている．また，“Syntax model” は表 4 が示すように，構文特徴のみを用いて，“Recognition Result” を誤り訂正した結果である．同

表 5　誤りタイプごとの評価
Table 5　Evaluation with each error type

	SUB	DEL	INS	COR	WER [%]
Recognition result	28,446	5,453	14,751	63,871	42.94
Syntax model	21,726	7,475	8,737	68,569	33.49
Semantic model (Baseline)	21,501	7,614	8,506	68,655	33.21
Proposed method	19,135	9,019	6,420	69,616	30.52

様に，“Semantic model (Baseline)” は構文特徴に加えて，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た．つまり，図 1 の Learning Semantic model から，Word re-ranking の処理を抜いて学習されたモデルである．“Proposed method” は，提案手法である図 1 に従って実験を行った結果である．

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．“Baseline” と比較すると，33.21 %から 30.52 %まで低下し，トータルで 2.69 ポイント改善した．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セマンティックスを利用できたことが示せた．また，構文特徴のみを用いた “Syntax model” と従来の “Semantic model” を比較すると，認識誤りに影響され，長距離言語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，有意な差は見られなかった．しかし，“Syntax model” と “Proposed method” を比較すると，WER が 2.97 ポイント改善しており，長距離文脈情報を効果的に用いることで有意な差が見られた．

しかし，図 3 が示すように，本実験では，1 段階目の訂正と 2 段階目の訂正において用いたモデルは，異なるデータセットから学習されている．一方で図 4 が示すように，両モデルとも異なるデータセットから，単語 N-gram を素性として学習している．これは単純に学習量が 2 倍になっているとも考えられる．従って，Semantic model を学習する際の素性を Long-distance context score のみにして，実験を行ってみる必要がある．

6. ま と め

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 段階目で構文特徴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，本提案手法と比較すると，単語誤り率は 33.21 %から 30.52 %まで，2.69 ポイント改善した．構文特徴のみを用いたモデルと比較するとトータルで 2.97 ポイント改善した．

5 章でも述べたように，今後の課題として，学習データにおける素性の選び方が挙げられる．また，CRF を改善した手法である Conditional Neural Fields [14] を利用することも考えたい．単語間類似度の計算方法については，現在利用している LSA 以外の手法も考えていきたい．

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”, 日本音響学会講演論文集（秋）, pp. 1–4, 2007.

[2] M. Goto, J. Ogata, K. Eto, “Podcastle: A Web2.0 Approach to Speech Recognition Research,” in *Proc. Interspeech2007*, pp. 2397–2400, 2007.

[3] J. Glass, T.J. Hazen, S. Cypher, I. Malioutov, D. Huynh, and R. Barzilay, “Recent progress in the mit spoken lecture processing project,” in *Proc. Interspeech2007*, pp. 2553–2556, 2007.

[4] 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.

[5] 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.

[6] Z. Zhou, J. Gao, F. K. Soong, and H. Meng, “A comparative study of discriminative methods for reranking lvcsr n-best hypotheses in domain adaptation and generalization,” in *Proc. ISCA*, pp. 1574–1577, 2006.

[7] A. Kobayashi, T. Oku, S. Homma, S. Sato, T. Imai, and T. Takagi, “Discriminative rescoring based on minimization of word errors for transcribing broadcast news,” in *Proc. ISCA*, pp. 1574–1577, 2008.

[8] 中谷良平, 滝口哲也, 有木康雄, “Confusion Network を用いた CRF による音声認識誤り訂正”, 第 5 回音声ドキュメント処理ワークショップ, 2011.

[9] 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.

[10] 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.

[11] Thomas Landauer, Peter W. Foltz, Darrell Laham, “Introduction to Latent Semantic Analysis,” in *Discourse Processing*, pp. 259-284, 1998.

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

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

[14] J. Xu J. Peng, L. Bo, “Conditional neural fields,” in *NIPS2009*, pp. 1419–1427, 2009.