Normalized Web Distanceを用いた音声認識誤りの訂正法

エンフボロル ビャムバヒシグ[1]　田中 克幸[2]　滝口 哲也[3]　有木 康雄[3]

概要：本稿では，従来のConfusion Networkに基づく音声認識誤り訂正で, ヌル遷移による短距離訂正の劣化と, 文脈スコアを計算するためのコーパスの必要性という問題点を指摘し, これらの問題点を解決するために以下の２つのアプローチで認識誤りの削減をねらう．１つ目は, 離れた単語も視野に入れ訂正する長距離文脈スコアとしてNormalized Web Distanceを用いる．Normalized Web Distanceは学習コーパスとして, World Wide Web, 検索エンジンなど様々なデータベースを利用することができるため, コーパスを用意する必要がなく, 計算も簡単にできるというメリットがある．２つ目は, 短距離訂正で有効であるN-gram学習において, 悪影響を及ぼすヌル遷移をテストデータから効率的に削除することにより, その効果を改善することで音声認識精度を改善する．実験の結果, 提案手法により単語誤り率を削減できることを確認した．

1.　はじめに

近年, 自動音声応答サービスやスマートフォン音声エージェントや自動字幕システム, さらに音声文字入力など音声認識システムの利用が普及し幅広く研究されている [1]. 現在までの音声研究の結果，音声認識は目覚ましい発展を遂げてきた．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 [2]．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．

しかしながら, 雑音環境や話者の発音や声質あるいは音声認識システムの語彙数など様々の要因により音声認識誤りが起きてしまう [3]. 現在の音声認識では, 言語モデルと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，音声認識誤りを避けることは難しい．そのため，音声認識精度の改善が望まれている．今まで, 音声認識精度の改善を図るため, 音声認識誤り訂正の手法が数多く提案されている．その中で, 識別モデルを採用し言語的に自然か不自然かということを学習した上で, 誤り訂正を行う手法がある．識別モデルの学習において重要な要素の一つは素性である．

従来，識別モデルにおける音声認識誤り訂正の素性としてunigram, bigram, trigramなどの単語N-gramや認識信頼などを用いることが多い．しかし, これだけでは付近の数単語のみとの意味的類似性は見れるが, 離れている単語間の類似性を見ることができない．また, 認識結果に誤りやConfusion Networkにおけるヌル遷移などが多く存在する際には短距離での学習・訂正さえ難しい場合がある．先行研究に離れた単語間の類似性を考慮し訂正する手法が提案されているが, 学習コーパスの用意の必要性やコーパス拡張に対する計算量問題などがある [4].

本稿では, これらの問題点を解決するために, 以下の２つのアプローチで認識誤りの削減をねらう．１つ目は, 離れた単語も視野に入れ訂正する長距離文脈スコアとしてNormalized Web Distance(NWD)[5]を用いることである. NWDは学習コーパスとして, World Wide Web, 検索エンジンなど様々なデータベースを利用することができるため, コーパスを用意する必要がなく, 計算も簡単にできるというメリットがある．２つ目は, 短距離訂正で有効であるN-gram学習において, 悪影響を及ぼすヌル遷移をテストデータから効率的に削除することにより, その効果を改善することである．まず, ヌル遷移を少しでも正確に検出・学習し次の段階で削除するため, ヌル遷移を残して学習した「ヌル遷移ありの検出モデル」を用いて一回目の訂正を行う．次に, 一回目の訂正結果から真と判断されたヌル遷移を削除し, その後, ヌル遷移を削除して学習した「ヌル遷移なしの検出モデル」を用いて２回目の訂正を行うことに

---

[1] 神戸大学大学院システム情報学研究科，兵庫県
Graduate School of System Informatics, Kobe University, Rokkodai 1–1, Nada-ku, Hyogo, 657–8501 Japan

[2] 神戸大学大学院経済学研究科，兵庫県
Graduate School of Economics, Kobe University, Rokkodai 2–1, Nada-ku, Hyogo, 657–8501 Japan

[3] 神戸大学自然科学系先端融合研究環，兵庫県
Organization of Advanced Science and Technology, Kobe University, Rokkodai 1–1, Nada-ku, Hyogo, 657–8501 Japan

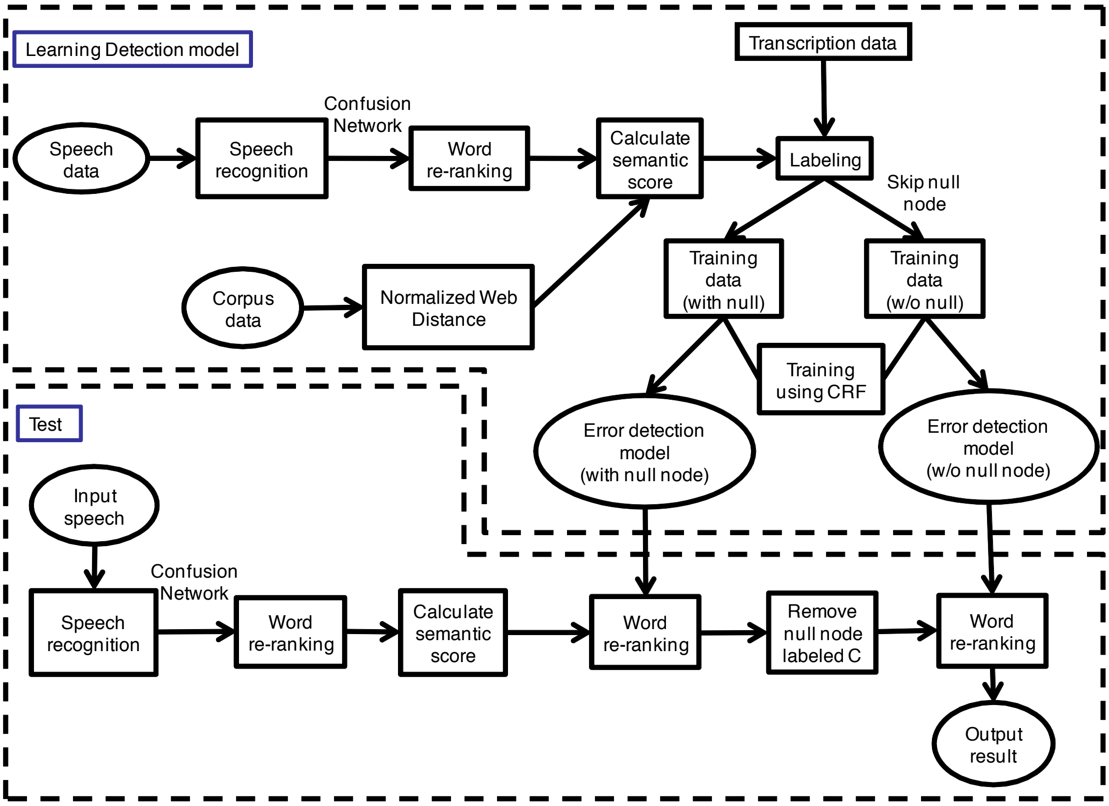

図 1　NWD 文脈スコアを用いた提案手法の流れ

Fig. 1　Flow of learning and detection process based on the NWD semantic score

より音声認識精度を向上させる．以降 2 章では，提案手法の流れについてついて述べる．3 章では Normalized Web Distance を用いた長距離文脈スコアついて，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 章で本稿の成果をまとめ，今後の課題について議論する.

2.　提案手法の流れ

図 1 に提案手法全体の流れを示している．上側の点線で囲まれた Learning detection model(学習) プロセスは，Normalized Web Distance （NWD) [5] による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．学習ではまず音声データを音声認識器に入力し，Confusion Network を出力する．出力された Confusion Network の特定の単語に NWD による長距離文脈スコア付与し，対応する書き起こしデータを元に正誤ラベルを貼る．正誤ラベル情報と unigram, bigram, trigram, Confusion network 上の存在確率, NWD 長距離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．

ここで，学習するのはトレーニングデータからヌル遷移を削除して学習したヌル遷移なしの検出モデルとヌル遷移ありの検出モデル２種類のモデルである．

テストデータ

うち	
-	
は	
どこ	
まで	
広がっ	
-	
-	
れ	
-	
-	
いる	
だろ	
う	

ヌル遷移ありの検出モデルで訂正後

宇宙	正
-	正
は	正
どこ	正
まで	正
広がっ	正
-	誤
-	正
れ	誤
-	正
-	正
いる	正
だろ	正
う	正

正ラベルのヌル遷移を削除後

宇宙	正
は	正
どこ	正
まで	正
広がっ	正
-	誤
れ	誤
いる	正
だろ	正
う	正

ヌル遷移なしの検出モデルで訂正後

宇宙	正
は	正
どこ	正
まで	正
広がっ	正
-	誤
て	正
いる	正
だろ	正
う	正

図 2　誤り訂正の流れ

Fig. 2　Flow of error correction process

図 1 下部のテストプロセスは誤り訂正を行う評価実験のプロセスである．ここでは学習プロセスと同様, 入力音声に対して Confusion Network を生成し，特定の単語に NWD による長距離文脈スコアを付与する．そして, 学習したヌル遷移あり (NWD context model w/ null) の検出モデルと

ヌル遷移なし検出モデル (NWD context model w/o null) によって 2 回誤り訂正を行う. まずは, ヌル遷移ありの検出モデルを用いて誤り訂正を行う. ここで, ヌル遷移ありのモデルを用いるのは, テストデータに含まれるヌル遷移を少しでも正確に検出し, 次の段階で削除するためである. その後, 訂正結果の内, 正しいと判断されたヌル遷移を削除した上で, ヌル遷移なしのモデルを用いてもう一回誤り訂正を行う.

このように 2 回の訂正を行うことによって以下の効果が期待される. ヌル遷移や認識誤りが多い場合は, ヌル遷移に影響を受けにくい長距離文脈情報がより力を発揮し認識誤りが訂正されると共に, ヌル遷移削除のための情報 (真と判断されたヌル遷移) を得ることができる. この一回目の訂正で, 短距離訂正に悪影響を及ぼした認識誤りとヌル遷移が削減されるため, 二回目の訂正では短距離の訂正が力を発揮し長距離では訂正しにくかった箇所が訂正できる. 図 2 に誤り訂正の例を示す. 青色で示したのは訂正された語, 赤色で示したのは削除対象となるヌル遷移である.

3. 長距離文脈情報

3.1 Normalized Web Distance

Normalized Web distance (NWD) は Google 検索を用いた Normalized Google Distance(NGD) [6] から由来した手法であるが, Google 検索だけではなく他の検索エンジンや World Wide Web やデータベースなど多様に適用できるよう提案したものである. NWD は意味の関わりの強さを測る尺度を表す事ができる手法として提唱されており, 正規化情報距離 (Normalized Information Distance) を近似したものである. 正規化情報距離はある表現 x とある表現 y の間の距離を示す指標であり, 下記の式で表される.

$$NID(x,y) = \frac{K(x,y) - \min(K(x),K(y))}{\max(K(x),K(y))} \tag{1}$$

ここで, $K(x)$, $K(y)$, $K(x,y)$ はそれぞれ文字列 x, y, x かつ y のコルモゴロフ複雑性 [7] を表している. コルモゴロフ複雑性 $K(x)$ は, 万能チューリングマシンなどの万能計算機において文字列 x を出力するために必要となるプログラム記述の長さ (bits) の最小値のことであり, 下記の式で表される.

$$K_U(x) = \min_{p:U(p)=x} l(p) \tag{2}$$

だたし, U は万能計算機, p はプログラムの記述, $l(p)$ は記述の長さである. 簡単に言うと, コルモゴロフ複雑性は, そのデータを記述するためにはどれだけの長さの記述が必要かということを示している. しかしながら, コルモゴロフ複雑性を任意の文字列において計算することは原理的に不可能である. 仮に, $K(x)$ を求めることができるプログラム $Q(x)$ があったと仮定し, 以下のプログラム $P(n)$ を解くとしよう.

$P(n)$: すべての $x \in \{0,1\}^n$ において, もし $K(x) \geq n$ であれば出力する.

上記のプログラム結果としてそれぞれ入力 n に対応する $x_1, x_2, ..., x_n$ が出力されたとする. ここで, $Q(x)$ の記述が一定であることと n の記述に必要なビット数が $2\log n$ であることから $K(x_n) \leq c + 2\log n$ であることがわかる. しかし, これはプログラム $P(x)$ 内の条件 $K(x_n) > n$ と矛盾している. したがって, 背理法によりコルモゴロフ複雑性は任意の文字列において計算不可能であることが証明される.

このため, 正規化情報距離を求めることも不可能ということになる. したがって, これを解決するために, コルモゴロフ複雑性の代わりに, 検索エンジンで検索し得られたページ数 (ヒット数) で近似することで計算できるようにしたのが NWD である. ある表現 x と y の間の Normalized Web Distance は以下のように求まる.

$$NWD(x,y) = \frac{\max(\log f(x), \log f(y)) - \log f(x,y)}{\log N - \min(\log f(x), \log f(y))} \tag{3}$$

ここでそれぞれ, $f(x)$ は表現 x を Google など検索エンジンで検索した時のヒット数, $f(y)$ は表現 y を検索した時のヒット数 , $f(x,y)$ は表現 y かつ表現 x を検索した時のヒット数, N は検索エンジンがインデックスした総ページ数である.

一般的に Normalilzed Web Distance は 0 から無限大までの値を取るが 0 ∼1 までの値が多い. 表現 x と表現 y が常に一緒に起きるまたは同値の場合 NWD はゼロとなる. またそれぞれ起きるが, 一緒に起きることがない場合は NWD が無限大となる. 表現 x と表現 y のどちらが起きない場合は無限大/無限大で 1 となる.

すなわち, $\mathrm{NWD}(x,y) = 0$ とは表現 x と表現 y が意味的に一致し, $\mathrm{NWD}(x,y) = \infty$ であれば意味の関わりが全くないことになる.

3.2 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 3 のように，意味的に不自然な単語が存在する場合に, そ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．

本稿では, どの単語と共起しても不自然でない「が」や「ます」といった機能語に対しては文脈スコアを付けず, 名詞, 動詞, 形容詞のみに与える. 長距離文脈スコアとして上記で紹介した Normalized Web Distance を用いる. また, NWD が無限大の場合, 計算簡略のため 1 とした. 音声認識結果に出現した内容語 w の長距離文脈スコア，$\mathrm{NWD}(w_i)$ は次のように計算する．

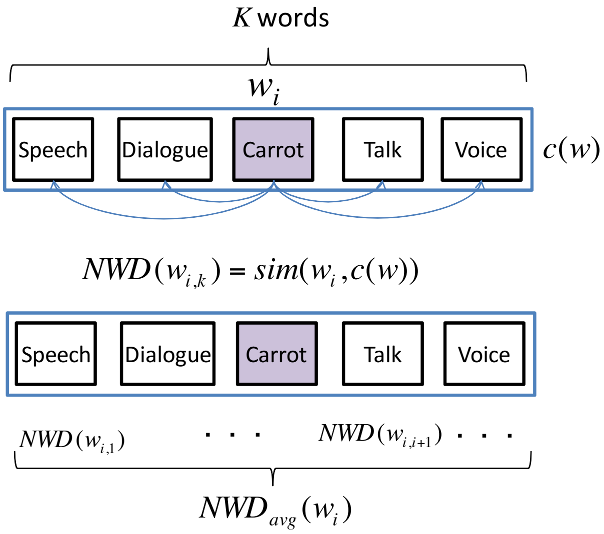

図 3 長距離文脈スコアの計算

Fig. 3 Calculation of semantic score

(1) w_i の周辺に現れる内容語を，図 3 のように文脈窓幅 K で集め，単語集合 $c(w)$ とする (w_i 自身は含まない).

(2) 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，NWD$(w_{i,k})$ とする.

$$NWD(w_{i,k}) = sim(w_i, c(w)) \qquad (4)$$

(3) NWD$(w_{i,k})$ から，平均 NWD$_{avg}(w_i)$ を求める.

$$NWD_{avg}(w_i) = \frac{1}{K}\sum_{k} NWD(w_{i,k}) \qquad (5)$$

(4) NWD$_{avg}(w_i)$ を w_i の長距離文脈スコアとする.

NWD$_{avg}(w_i)$ が小さ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，NWD$_{avg}(w_i)$ は全体的に大きくなる.

4. 音声誤り訂正

4.1 Conditional Random Fields

Conditional Random Field (CRF) [8] 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対して最適と推定するラベルを割り当てるのではなく，系列全体に対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できる．

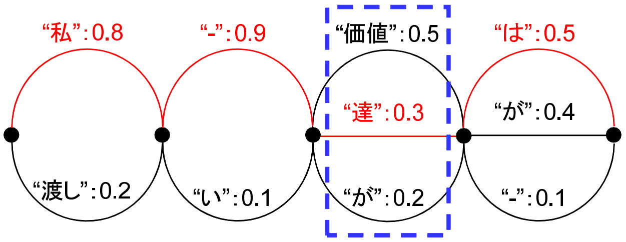

図 4 Confusion Network の例

Fig. 4 Example of Confusion Network

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．

$$P(y \mid x) = \frac{1}{Z(x)} \exp\Big(\sum_{a} \lambda_a f_a(y, x)\Big) \qquad (6)$$

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．

$$Z(x) = \sum_{y} \exp\Big(\sum_{a} \lambda_a f_a(y, x)\Big) \qquad (7)$$

パラメータ λ_a は，学習データ $(x_i, y_i)(1 \le i \le N)$ が与えられたとき，条件付確率分布 (6) の対数尤度，

$$\mathcal{L} = \sum_{i=1}^{N} \log P(y_i \mid x_i) \qquad (8)$$

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 [9] 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．$\hat{y}$，すなわち式 (9) は Viterbi アルゴリズムで効率的に解くことができる．

$$\hat{y} = \operatorname*{argmax}_{y} P(y \mid x) \qquad (9)$$

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 [10].

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．N-gramにおけるヌル遷移については，他の単語と同様にヌル遷移という単語として取り扱う．

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[11]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ので，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．第二候補単語列とは，単語を入れ替えた際のスコアが（最尤候補から数えて）二番目に大きくなる単語を並べたものである．以下同様に，第三候補単語列も，それぞれのConfusion Setにおいて信頼度が三番目となる単語を並べたものである．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．図1で示した通り誤り検出モデルの学習後, 以下のアルゴリズムに従って２回誤り訂正を行う．

一回目，「ヌル遷移ありの検出モデル」：

(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, テストデータとしCRFによる誤り検出を行って，正誤ラベルを付与する．
(3) 入力時系列順にテストデータを見ていく．正解と判定された語には何も操作を行わずに次の単語へ進む．誤りと判定された語は，対応するConfusion Setから次に存在確率が高い候補を選び出し，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せず, Confusion Setの中にヌル遷移があればそれを選択する．

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)
Number of mixtures	16
State	5 states and 3 loops

(5) Confusion Setの中に正解単語もヌル遷移も存在しなければ，存在確率の最も高い語を選択する．
(6) すべてのConfusion Setについて順番に3, 4, 5を繰り返す．

２回目,「ヌル遷移なしの検出モデル」：

(1) 一回目訂正後の出力からヌル遷移を選択削除したものをテストデータとし, 正誤ラベルも一回目訂正後のラベルを用いる．
(2) 以降は, 一回目の訂正手順２，３，４，５，６と同様である．

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

本稿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 [15]を用いる．市販のシステムとは異なり，様々な使用環境や目的に応じたシステムの構築が容易なため，研究目的などに多く使用されている．以下，システムに必要な音響モデルと言語モデルについて述べる．

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

表 2　N-gram のエントリー
Table 2　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　Number of data

	Detection model Training	Test
Number of lectures	150	301
Number of words	311,374	113,289

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．Julius は 2-pass 探索を行っており，前向き探索には bigram モデル，後ろ向き探索には trigram モデルを用いる．

また, 本稿では図 1 が示すように，NWD を計算するためのデータ (2 種類), 長距離文脈スコアを付与し誤り検出モデルを学習するためのデータ，評価データの計 4 つのデータセットを利用した．

NWD のコーパスとして比較のため次の２種類を用意した．一つ目は, CSJ の書き起こし文書，2,672 講演分のデータである．内容語として名詞，動詞，形容詞のみを検索対象とし，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．二つ目は, ウェブコーパスとして Yahoo!知恵袋データベースの回答数 5000 万件 2004 年 4 月－ 2009 年 4 月までの部分を利用した．図 3 における意味スコアを求める際の単語集合は前後 3 発話ずつの Confusion Network における存在確率最大の単語列とした．

誤り検出モデルの学習と，評価に用いたデータ数を表 3 に示す．誤り検出モデルの学習には NWD コーパスと異なる 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．

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つまり Confusion network の最尤候補 (CN-best) である．“N-gram model” は N-gram と Confusion network 上の信頼度を素性としたもの, “LSA context model (Baseline)” はそれに Latent semantic analysis(LSA)[12] による文脈スコアを加えた手法である．“NWD context model w/null” は文脈スコアとして Normalized Web Distance を用いたもの, “NWD context model w/o null” は上記と素性は一緒だが, 学習データからヌル遷移を削除したものである．それぞれ用いた素性を表 4 に示す．○ (使用)，× (未使用) を表示している．すべての手法は表 3 の学習データとテストデータを用いている．

“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．“Proposed method ” では検出モデルを２つ用いる．最初は “NWD context model w/null” で訂正した後, 評価データより正解判断されたヌル遷移を除く．その後に “NWD context model w/o null” を用いる．

表 5 で示すように，Normalized Web Distance を用いた誤り訂正はベースラインとなる LSA を用いた手法と比べると, 学習コーパスが Yahoo!知恵袋の場合と CSJ の場合で, それぞれ単語誤り率が 1.07 ポイントと 2.59 ポイント改善している．さらに, 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，単語誤り率が最も小さくなっている．“Baseline” と比較すると，32.77 %から 28.32 %まで低下し，トータルで 4.45 ポイント改善した．以下に, 提案手法による訂正例を示す．

- 正解文：これ は 驚き で あっ た
- 認識結果：れ は 驚き だっ た
- 1 回目訂正後： 驚き た
- 2 回目訂正後：これ 驚き あっ た

- 正解文：宇宙 は 今日 より 昨日 の 方 が 昨日 より おととい の 方 が 小さかっ た と いう こと だろ う
- 認識結果：っちゅう は 許容 より 昨日 の 方 が に の L　音 と の 方 が 近かっ た と いう こと を
- 1 回目訂正後： 宇宙 は 許容 より 昨日 の 方 よる おととい と の 方 が 近かっ た と いう こと
- 2 回目訂正後：宇宙 は 今日 より 昨日 の 方 今 よる おととい と の 方 が 近かっ た と いう こと が

2 番目の例では, 一回目の訂正において「っちゅう/宇宙」や「音と/おととい」など比較的トピックを持つ単語が正しく訂正されている．また, 両方の例から, ２回目の訂正では「れ/これ」や「許容より昨日/今日より昨日」など比較的短距離で訂正が多いことが分かる．

6.　おわりに

本稿では，N-gram による短距離と NWD による長距離言語情報を効率的に利用して音声認識誤りを自動訂正し，音声認識精度を向上させる手法を提案した．

NWD を用いた提案手法は, 従来手法である LSA による文脈を用いた誤り訂正手法と比べて単語間の類似度をよりよく表現し, 音声誤り訂正で有効であることを確認できた．また, 学習対象が Yahoo!知恵袋と CSJ といったコンセプトの異なるコーパスでも, 同等の訂正ができた．

表 4 検出モデルの学習に用いられた素性
Table 4 Features used in each model

	N-gram	Confidence score	LSA score	NWD score	Null node skip
Recognition result	×	×	×	×	×
N-gram model	○	○	×	×	×
LSA context model (Baseline)	○	○	○	×	×
NWD context model w/ null (1)	○	○	×	○	○
NWD context model w/o null (2)	○	○	×	○	×
Proposed method (1 + 2)	○ ○	○ ○	× ×	○ ○	○ ×

表 5 比較手法と提案手法の評価
Table 5 Evaluation of each method

	SUB	DEL	INS	COR	WER [%]
Recognition result	28,446	5,453	14,751	63,871	42.94
NWD context model w/o null	23,088	6,966	9,625	67,416	35.02
N-gram model	21,522	7,848	8,204	68,400	33.17
LSA context model (Baseline)	21,049	8,324	7,757	68,397	32.77
NWD context model w/null (Yahoo)	20,469	10,130	5,316	67,171	31.70
NWD context model w/null(CSJ)	18,073	11,524	4,597	67,873	30.18
Proposed method NWD w/null + NWD w/o null	15,118	13,534	3,431	68,794	28.32

また, 提案手法により従来手法と比べて単語誤り率が 32.77%から 28.32 まで, 4.45 ポイント改善した．これは提案手法により, 単語誤りやヌル遷移などが多い時に, ヌル遷移ありの検出モデルで訂正すると長距離文脈スコアで離れた単語の誤りが訂正され, その後ヌル遷移を効率的に削除し, ヌル遷移なしのモデルで訂正することにより短距離訂正の力が発揮されていると考えられる．

今後の課題として，誤り検出を行う際の識別器として，CRF を改善した手法である Conditional Neural Fields [16] を利用することも考えたい．また, 本稿では比較的トピックを持つ名詞, 動詞, 形容詞のみに NWD 文脈スコアを付与した．今後としては, 隣接する「さ」「せ」「て」などの助動詞にも文脈スコア付与することにより, 長距離だけでなく短距離の文脈性も見られるのではないかと思われる．

参考文献

[1] Jerome R.Bellegarda, “Large scale personal assistant technology deployment,” in *Proc. INTERSPEECH2013*, pp. 2029–2033, 2013.

[2] 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,” 日本音響学会研究発表会講演論文集 (秋), pp. 1–4, 2007.

[3] L. Deng and X. Huang, “Challenges in adopting speech recognition,” Communications of the ACM, vol. 47, No.1, pp. 60–75, 2004.

[4] Ryohei Nakatani, Tetsuya Takiguchi, Yasuo Ariki, “Two-step correction of speech recognition errors based on n-gram and long contextual information,” in *Proc. INTERSPEECH2013*, pp. 3747–3750, 2013.

[5] Cilibrasi, R.L., P.M.B. Vitanyi, “Normalized Web Distance and Word Similarity,” Handbook of Natural Language Processing, 2nd ed, pp. 293–314, 2010.

[6] Cilibrasi, R.L., P.M.B. Vitanyi, “The Google similarity distance,” IEEE Trans. Knowl. Data Eng., pp. 370–383, 2007.

[7] A.N. Kolmogorov. “Three approaches to the quantitative definition of information,” Problems Inform. Transmission, vol. 1, No.1, pp. 1–7, 1965.

[8] 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.

[9] J. Nocedal, “Updating Quasi-Newton Matrices with Limited Storage,” in *Mathematics of Computation*, pp. 773–782, 1980.

[10] Lidia Mangu, Eric Brillx, Andereas Stolcke, “Finding consensus in speech recognition: word error minimization and other applications of confusion networks,” *Computer Speech and Language*, vol. 14, pp. 373–400, 2000.

[11] 松本 智彦, 滝口 哲也, 有木 康雄, “識別的言語モデルに基づく Confusion Network 上での音声認識誤り訂正,” 日本音響学会研究発表会講演論文集 (春), pp. 175–176, 2010.

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

[13] 松本智彦, 滝口哲也, 有木康雄, “複数の言語情報を用いた CRF による音声認識誤りの検出,” 電子情報通信学会, 音声研究会, SP2008-127, pp. 7-12, 2009.

[14] 人間文化研究機構国立国語研究所, “日本語話し言葉コーパス (Corpus of Spontaneous Japanese),” http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/, 参照 2013-01-17.

[15] Julius development team, “大語彙連続音声認識エンジン Julius,” http://julius.sourceforge.jp/, 参照 2013-01-17.

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