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

SDPWS2014-03

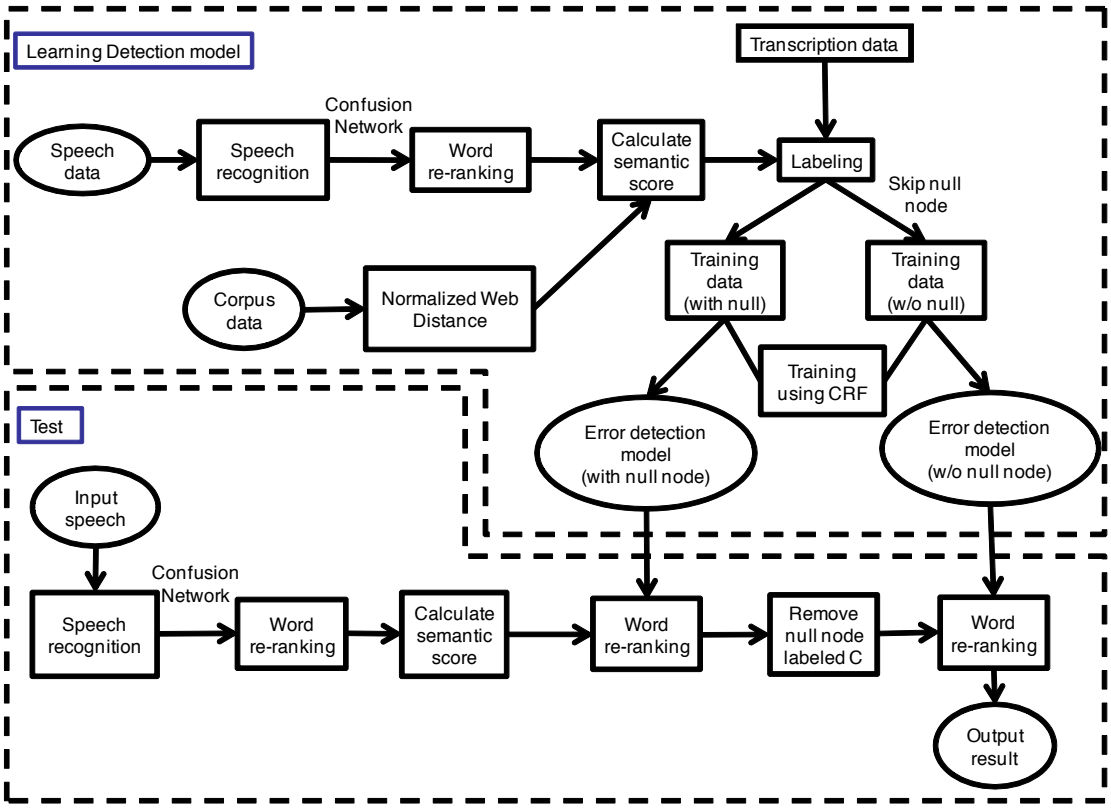

図 1　NWD 文脈スコアを用いた提案手法の流れ

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．

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．

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

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．

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

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．

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

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．

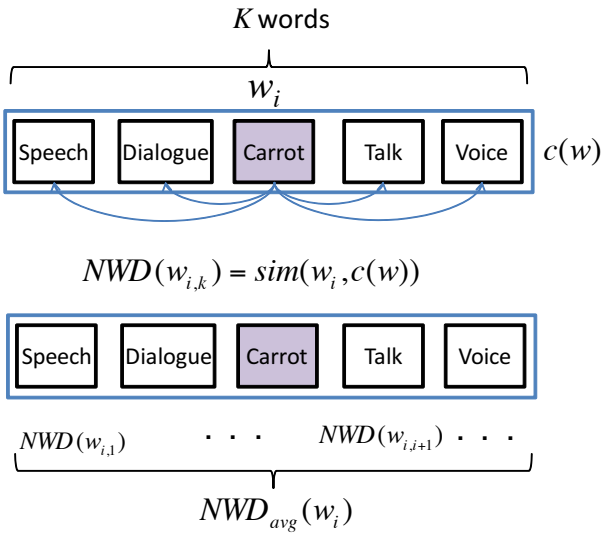

図 3 長距離文脈スコアの計算
Fig. 3 Calculation of semantic score

(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．

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

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

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

(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．

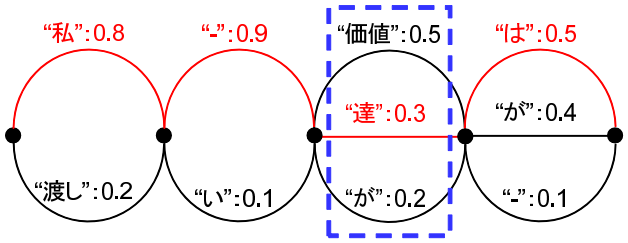

図 4 Confusion Network の例
Fig. 4 Example of Confusion Network

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．

$$P(y \mid x) = \frac{1}{Z(x)} \exp(\sum_a \lambda_a f_a(y, x)) \tag{6}$$

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．

$$Z(x) = \sum_y \exp(\sum_a \lambda_a f_a(y, x)) \tag{7}$$

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，

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

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．

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(y \mid x) \tag{9}$$

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