Normalized Similarity Distanceを用いた音声認識の誤り訂正法*

☆房安陽平, 滝口哲也, 有木康雄 (神戸大)

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, 2] 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 [3]．また会議などの話し言葉においても 90%程度の高度な認識が可能となっている [4]．しかし, 現在の音声認識では完全に音声認識誤りを避けることは難しい．よって単純に音声認識システムの精度を上げること以外のアプローチで Word Error Rate (WER) を低くすることが求められる．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 [5].

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, 単語間の情報距離である Normalized Similarity Distance として Normalized Relevance Distance (NRD) [6] を利用した長距離文脈情報を用いることである．NRD は単語間の類似度, 及び単語と検索に利用した文書 (ページ) との関係性を考慮に入れた情報距離であるため誤り訂正の精度向上が期待できる．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．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長距離文脈情報による誤り検出で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より音声認識精度を向上させる．

2 長距離文脈情報

2.1 Normalized Relevance Distance

Normalized Relevance Distance は単語間の意味の関わりの強さを測る尺度として用いることができる手法として提唱されており, Normalized Web Distance (NWD) [7] を変形したものである.NWD も同様に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．正規化情報距離はその定義の中にコルモゴロフ複雑性を含んでいる．コルモゴロフ複雑性の計算は原理的に不可能である．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 であり, 検索エンジンを利用して得られた tf-idf で近似することで計算できるようにしたのが NRD である．ある表現 x と y の間の Normalized Relevance Distance は以下のように求まる．

$$NRD(x,y) = \frac{\max(\log f_{NRD}(x), \log f_{NRD}(y)) - \log f_{NRD}(x,y)}{\log N - \min(\log f_{NRD}(x), \log f_{NRD}(y))} \tag{1}$$

$$f_{NRD}(x) = \sum_{d \in D} tfidf_{norm}(x,d) \tag{2}$$

$$f_{NRD}(x,y) = \sum_{d \in D} tfidf_{norm}(x,d) \cdot tfidf_{norm}(y,d) \tag{3}$$

$f_{NRD}(x)$ は単語 x とある文書 d に対して tf-idf を計算し, 単語 x で検索した時にヒットした全ての文書数 D に対して同様の処理を行い, その総和を計算した

*WORD-ERROR CORRECTION USING NORMALIZED SIMILARITY DISTANCE FOR CONTINUOUS SPEECH RECOGNITION(Kobe University)

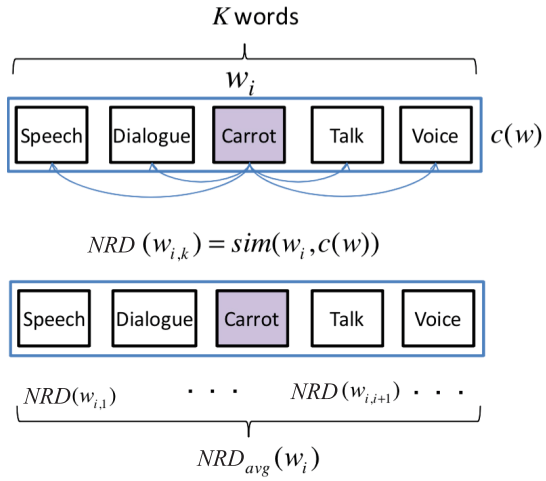

Fig. 1　長距離文脈スコアの計算

ものである．$f_{NRD}(x, y)$ は単語 x, 単語 y とある文書 d に対してそれぞれ個別に tf-idf を求めたものの積を，単語 x かつ単語 y で検査した時にヒットした全ての文書数 D に対して同様の処理を行い，その総和を計算したものである．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．

2.2　長距離文脈スコアの計算

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．長距離文脈スコアを計算する情報距離として上記で紹介した Normalized Relevance Distance を用いる．音声認識結果に出現した内容語 w の長距離文脈スコア，$\mathrm{NRD}_{avg}(w_i)$ は次のように計算する．

1. w_i の周辺に現れる内容語を，図 1 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，$\mathrm{NRD}(w_{i,k})$ とする．

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

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

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

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

$\mathrm{NRD}_{avg}(w_i)$ 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，$\mathrm{NRD}_{avg}(w_i)$ は全体的に大きくなる．

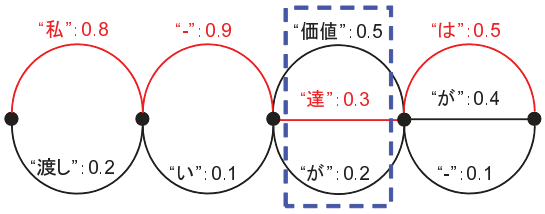

Fig. 2　Confusion Network の例

3　音声認識の誤り訂正

3.1　Conditional Random Fields

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，Conditional Random Fields(CRF) [8] 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．

3.2　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 [9] を用いる．

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 2 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 2 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 2 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．

3.3　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．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．
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．

二回目，「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．

4　評価実験

4.1　実験条件

コーパスは日本語話し言葉コーパス（CSJ）[10] を用いた．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．

Table 2　N-gram のエントリー

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

Table 3　学習, 評価データ数

	Training	Test
Number of lectures	450	100
Number of words	508,299	29,162

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 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．

また, 本稿では NRD を計算するためのデータ, 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 3 つのデータセットを利用した．

NRD の計算用コーパスとしては 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．

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 NRD コーパスと異なる 450 講演分の音声データ，評価には学習データを含まない 100 講演分の音声データをそれぞれ用いた．

また, NRD との比較のために NWD を用いた誤り訂正も同様に行った．

4.2　実験結果

実験結果とそれぞれに用いた素性を表 1 に示す．“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．“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．NRD の場合も同様に対応している．また, ○ (使用)，×

Table 1　各手法で用いた素性及び誤り訂正実行後の単語誤り率 (Word-Error Rate)

	N-gram	Confidence score	NWD	NRD	Null node skip	WER [%]
Recognition result (Baseline)	×	×	×	×	×	43.52
N-gram model	○	○	×	×	×	33.45
NWD context model w/ null (1)	○	○	○	×	○	35.49
NWD context model w/o null (2)	○	○	○	×	×	30.79
NWD (1 + 2)	○ ○	○ ○	○ ○	× ×	○ ×	29.19
NRD context model w/ null (1')	○	○	×	○	○	37.42
NRD context model w/o null (2')	○	○	×	○	×	34.12
Proposed method (1' + 2')	○ ○	○ ○	× ×	○ ○	○ ×	28.09

(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．

表 1 で示すように，Normalized Relevance Distance を用いた誤り訂正は NWD を用いた手法と比べると, 単語誤り率が 1.09 ポイント改善している．また,N-gram モデルと比べると NRD, NWD どちらの長距離文脈スコアを用いた場合も大幅に改善が見られた．

5　おわりに

本稿では，*N*-gram による短距離と NRD 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．

NRD を用いた提案手法は, 従来手法である NWD 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,N-garm のみを用いた手法と比べると, 長距離文脈スコアを用いた WER の改善率は大幅に上昇した．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,「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] J. R.Bellegarda, "Large scale personal assistant technology deployment," in Proc. INTERSPEECH2013, pp.2029-2033, 2013.

[2] F. Burkhardt and H. U. Nageli, "Voice search in mobile applications and the use of linked open data," in Proc. INTERSPEECH2013, pp. 2059-2061, 2013.

[3] 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.

[4] 河原達也, "話し言葉の音声認識の進展—議会の会議録作成から講演・講義の字幕付与へ—," メディア教育研究, vol. 9, no. 1, pp. S1–S8, 2012.

[5] R. Nakatani *et al.*, "Two-step correction of speech recognition errors based on n-gram and long contextual information," in Interspeech, pp. 3747–3750, 2013.

[6] C. Schaefer *et al.*, "Normalized relevance distance a stable metric for computing semantic relatedness over reference corpora," ECAI 2014, vol. 263, pp. 789 – 794, 20.

[7] Cilibrasi *et al.*, "Normalized web distance and word similarity," Handbook of Natural Language Processing, vol. 2, pp. 293–314, 2010.

[8] J. D. Lafferty *et al.*, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in ICML, 2001.

[9] L. Mangu *et al.*, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," Computer Speech and Language, vol. 14, pp. 373–400, 2000.

[10] 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/.