301-4in

Normalized Web Distanceを用いた音声認識誤りの訂正法

Error correction of automatic speech recognition based on Normalized Web Distance

エンフボロル ビャムバヒシグ[*1] Byambakhishig Enkhbolor
田中 克幸[*2] Katsuyuki Tanaka
相原 龍[*1] Ryo Aihara
滝口 哲也[*3] Tetsuya Takiguchi
有木 康雄[*3] Yasuo Ariki

[*1]神戸大学大学院システム情報学研究科 Graduate School of System Informatics
[*2]神戸大学大学院経済学研究科 Graduate School of Economics
[*3]神戸大学自然学系先端融合研究環 Organization of Advanced Science and Tecnology

In this paper we focus on the problems in ASR error correction method based on Confusion Networks, which are degradation of N-gram correction due to the null transitions and the availability of corpus in terms of calculating semantic score. In attempt to solve these problems, first, we employ Normalized Web Distance as a measure for semantic similarity between words which are distant from each other. The advantage of Normalized Web Distance is that it uses WWW, Search engines and transcripts as a database for learning, which can solve the problem of availability of corpus and the computational complexity. Secondly, we delete the null transitions from the test data for N-grams to learn and correct effectively. Experimental results show that the correction of speech recognition using our proposed method can reduce word errors.

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. 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. 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. 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 [1].

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)[2] 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より音声認識精度を向上させる.

連絡先: 相原 龍, 神戸大学大学院システム情報学研究科, 神戸市灘区六甲台町１－１, aihara@me.cs.scitec.kobe-u.ac.jp

2. 長距離文脈情報

2.1 Normalized Web Distance

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.

$$NWD(x,y) = \frac{\max(\log f(x), \log f(y)) - \log f(x,y)}{\log N - \min(\log f(x), \log f(y))} \quad (1)$$

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

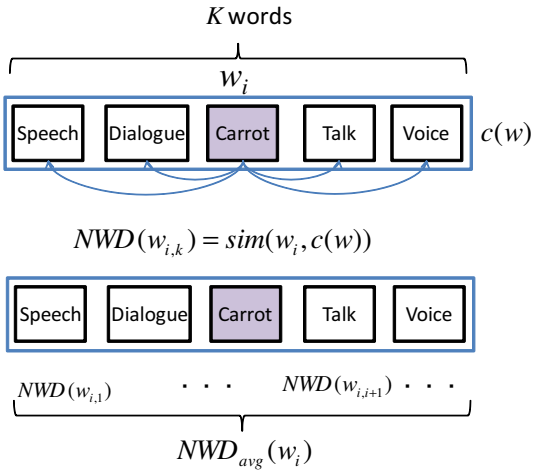

図 1: 長距離文脈スコアの計算

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．

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，NWD(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，NWD($w_{i,k}$) とする．

$$NWD(w_{i,k}) = sim(w_i, c(w)) \tag{2}$$

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

$$NWD_{avg}(w_i) = \frac{1}{K} \sum_k NWD(w_{i,k}) \tag{3}$$

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．

3. 音声誤り訂正

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) [3] 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．

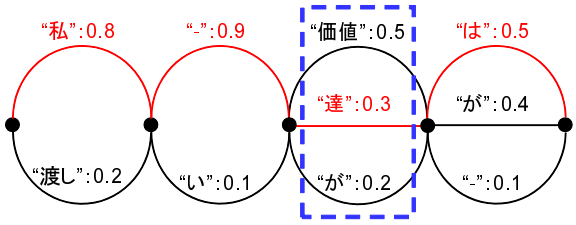

図 2: Confusion Network の例

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．

$$P(y \mid x) = \frac{1}{Z(x)} \exp(\sum_a \lambda_a f_a(y, x)) \tag{4}$$

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．

$$Z(x) = \sum_y \exp(\sum_a \lambda_a f_a(y, x)) \tag{5}$$

パラメータ λ_a は，学習データ $(x_i, y_i)(1 \leq i \leq N)$ が与えられたとき，条件付確率分布 (4) の対数尤度，

$$\mathcal{L} = \sum_{i=1}^{N} \log P(y_i \mid x_i) \tag{6}$$

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 (7) は Viterbi アルゴリズムで効率的に解くことができる．

$$\hat{y} = \underset{y}{\text{argmax}}\, P(y \mid x) \tag{7}$$

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 [4] 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

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 [5] 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．データは日本語話し言葉コーパス（CSJ） [6] 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．

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．N-gram エントリーは表 1 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．

表 1: N-gram のエントリー

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

表 2: 学習, 評価データ数

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

また, 本稿では 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 1 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 2 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．

4.2 実験結果

表 4 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)[7] 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 3 に示す．○ (使用)，× (未使用) を表示している．すべての手法は表 2 の学習データとテストデータを用いている．“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 4 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，

表 3: 検出モデルの学習に用いられた素性

	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 4: 比較手法と提案手法の評価

	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

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

5. 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.

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, 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] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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