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*

◎松本智彦, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

大語彙連続音声認識において，仮説集合をリランキングすることで認識精度を向上させる，識別的言語モデルが提案されている [1][2]．この手法は，音声認識器の出力した仮説集合と，それに対応する正解単語列から，認識誤りを特徴づける N-gram を学習し，リランキングに用いる．本稿では，Confusion Network 上でこの手法を適用し，誤り訂正を行う．日本語話し言葉コーパスによる実験から，単語誤り率の改善が見られたので報告する．

2 Confusion Network 上でのモデル学習と誤り訂正

2.1 Confusion Network

Confusion Network(CN)[4] とは，情報をコンパクトにまとめた仮説集合の表現方法である．CN の具体例として,“ 私達は ”という発話を入力したときのものを図 1 に示す．破線で囲まれた遷移の集合は，時間的な競合単語を表しており，この集合を Confusion Set(CS) と呼ぶ.“ - ”は単語が存在しないヌル遷移を表している．また，CN 上の各単語は，CS における存在確率 (信頼度) を持っている．

2.2 識別的言語モデル

従来の誤り訂正は，N-best のような仮説単語列集合から，最も誤りの少ない単語列を選択することを考えるため，柔軟な誤り訂正を行うことができない．そこで本稿では，CN を利用することで単語単位での誤り訂正を行う．

CN 上の単語 w_i の信頼度を $RecScore(w_i)$，音声認識スコア重みを λ，単語 w_i を特徴づける N-gram 素性ベクトルを $\Phi(w_i)$，素性の重みベクトルを α としたとき，各 CS において以下の式により単語を選択していく．

$$w^* = \underset{w_i \in CS}{\operatorname{argmax}} \{\lambda \cdot RecScore(w_i) + \alpha \cdot \Phi(w_i)\}. \quad (1)$$

学習はパラメータ α を推定する問題となるが，本稿ではパーセプトロンアルゴリズム [1] を用いる．学習の手順として，まず，パラメータ α を零ベクトルで初期化する．次に，以下の 2 式を学習データ中のすべての CS に対して繰り返し適用し，α を更新していく．

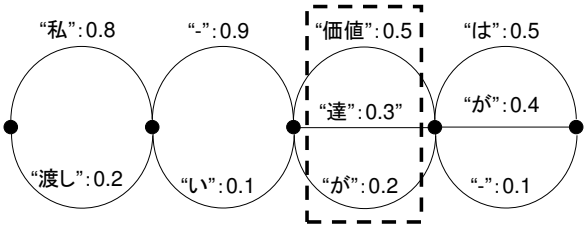

Fig. 1 Confusion Network の具体例

$$w^* = \underset{w_i \in CS}{\operatorname{argmax}} \{\alpha \cdot \Phi(w_i)\}. \quad (2)$$

$$\alpha = \alpha + \Phi(w^{ref}) - \Phi(w^*). \quad (3)$$

w^{ref} は，その CS における正解単語である．w^{ref} が存在しない CS では学習を行わないこととする．(3) 式により，w^{ref} と w^* が一致しない場合，正解単語の持つ素性の重みは正方向へ，誤り単語の持つ素性の重みは負方向に更新される．

ここで，素性ベクトル $\Phi(w_i)$ について説明する．$\Phi(w_i)$ は，単語 w_i がある N-gram を構成していれば 1，そうでなければ 0 となる，2 値の素性関数を要素とするベクトルである．本稿では，付近の CS における信頼度最大の単語から素性ベクトルを求めることにする．この際，信頼度最大の単語がヌル遷移であった場合，その CS はスキップする．例として，図 1 において“ 価値 ”という単語の素性ベクトルを求めることを考える．この際,“ 価値 ”の 1 つ前の CS において信頼度最大の単語であるヌル遷移はスキップすることで,“ 私/価値 ”や“ 私/価値/は ”という N-gram を生成する．

3 評価実験

3.1 実験条件

日本語話し言葉コーパス (CSJ) による評価実験を行った．学習と評価に用いたデータは表 1 のようになっている．音声認識器には Julius[5] を用い，発話ごとに CN を出力した．音声認識では，言語モデルに CSJ から学習した trigram を用いた．

提案する識別的言語モデルの学習には，CN の各単語に正誤のラベルが必要となるが，CN と正解文書と

*Speech Recognition Error Correction using Confusion Network Based on Discriminative Language Model, by Tomohiko Matsumoto, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

Table 1　実験に用いたデータ

	学習	評価
講演数	150	10
発話数	54,887	2,891
単語数	497,755	22,649
CS 数	883,977	45,244

Table 2　素性数と誤りの種類別評価

	素性数	SUB	DEL	INS	COR	WER
CN-oracle	-	1,613	2,213	579	18,807	19.62
CN-best	-	4,750	2,120	1,685	15,768	37.79
no-skip	770,305	4,197	2,618	1,139	15,823	35.14
N-gram	1,195,636	4,198	2,395	1,278	16,045	34.77
semantic	1,240,618	4,196	2,388	1,253	16,054	34.62

の間で DP マッチングをすることで自動でラベリングを行った．素性には，表層単語の 1-gram，2-gram，3-gram，さらに Latent Semantic Analysis に基づく意味スコア [3] も用いた．パーセプトロン学習の繰り返し回数はすべて 10 とし，繰り返しごとのパラメータの平均値を用いた．

評価値としては単語誤り率 (WER) を用い比較を行った．

3.2　実験結果

図 2 は音声認識スコア重み λ を変化させたときの WER，表 2 は各モデルにおける素性数と，WER が最も低かったときの誤りの種類別個数である．CN-oracle は各 CS において常に正解の単語を選択したもので，上限値と言える．CN-best は各 CS において信頼度が最大の単語を選択したもので，これをベースラインとする．no-skip はヌル遷移をスキップせずに N-gram 素性を求めたモデル，N-gram はヌル遷移をスキップしたモデル，semantic は意味スコアを素性として追加したモデルである．

どのモデルを用いた場合でも CN-best と比較して WER が改善しているが，ヌル遷移をスキップしたモデルのほうが，スキップしていないモデルに比べて 0.37 ポイント WER が低い．これは，前後にヌル遷移があるという情報は，正誤の識別にあまり影響を及ぼさないからであると考えられる．また，意味スコアを追加することで，さらに 0.15 ポイント改善し，CN-best と比較すると 3.17 ポイントの改善が見られた．

誤りの種類別に見ると，置換誤りと挿入誤りは大きく減少しているが，削除誤りは増加してしまっている．これは，認識誤りの多くをヌル遷移で置換することで，削除していることを表していると考えられる．ヌル遷移は通常の単語と比べて識別が困難であり，また，正解として出現することが多い．そのため，ヌル遷移を選択しやすいようなモデルが学習されている可能性がある．本稿では正解単語が存在する CS でのみ学習を行ったが，正誤のラベリングを行う際，正解単語が存在しない CS に正解単語を付与しておくことで，モデルの高精度化が図れるのではないかと考えられる．

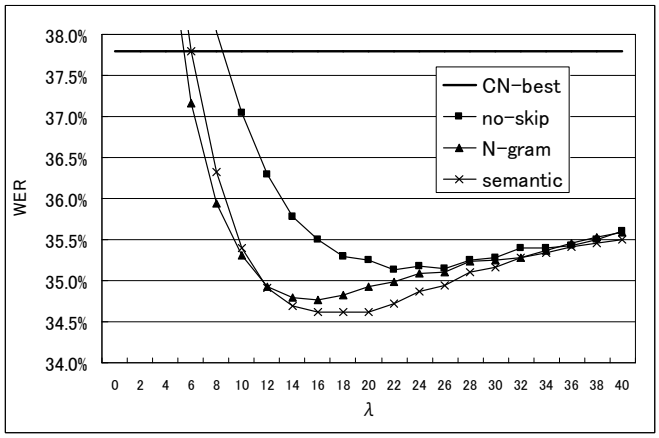

Fig. 2　λ による単語誤り率の変化

4　おわりに

本稿では，識別的言語モデルにより CN の仮説をリランキングし，音声認識精度の改善を試みた．日本語話し言葉コーパスによる評価実験で，提案手法による WER の改善が確認された．今後の課題として，N-best やラティスを利用して誤り訂正を行った場合との比較を行う必要がある．また，素性として品詞情報などを利用することや，最大エントロピー法による学習法を取り入れることで，より高精度なパラメータ推定を行うことも考えたい．

参考文献

[1] B. Roark, et al, “Discriminative language modeling withconditional random fields and the perceptron algorithm”, ACL, pp.47-54, 2004.

[2] 大庭, 他, “単語誤り率を考慮した誤り訂正モデル学習とその効果に関する分析”, 音講論 (春), pp.127-128, 2008.

[3] 松本, 他, “複数の言語情報を用いた CRF による音声認識誤りの検出”, 音講論 (春), pp.227-228, 2009.

[4] L. Mangu, et al, “Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks”, Computer Speech and Language, pp373-400, 2000.

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