CRF を用いた音声認識誤り訂正における素性の検討*

☆中谷良平, 滝口哲也, 有木康雄 (神戸大)

1　はじめに

我々は，大語彙連続音声認識において，Conditional Random Fields (CRF) [1] を用いて認識結果中の誤りを訂正する手法を提案してきた [2]．素性として，長距離言語情報などを用いたが，あまり大きな効果が得られなかった．そのため，本稿では，長距離言語情報を他の情報と組み合わせ，新たな素性として誤り訂正に用いる．その結果，長距離言語情報を単独で用いた場合と比較して，単語誤り率の改善が見られたので報告する．

2　モデル学習と誤り訂正

音声認識誤り訂正は CRF による誤り検出を応用して行う．認識単語列に対して誤り検出を行い，誤認識と識別された単語を Confusion Network [3] 中の競合候補と置き換えることで，誤り訂正を実現する．

2.1　CRF による誤り検出モデルの学習

本稿では，誤り検出モデルを 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．

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

ここで f_a は素性，λ_a は素性関数に対する重みである．$Z(x)$ は分配関数で，次式で与えられる．

$$Z(x) = \sum_{y} \exp(\sum_{a} \lambda_a f_a(y, x)) \qquad (2)$$

パラメータ λ_a は，学習データが与えられたとき，条件付確率分布 (1) の対数尤度を最大にするように学習される．

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．

2.2　長距離言語情報

本稿では，長距離言語情報として意味スコアを用いる．意味スコアとは，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不自然でないかという情報のことである．例えば Fig. 1 のように，「音声」,「会話」,「話者」,「対話」などが含まれる単語列の中に,「大根」という単語が含まれる場合，明らかに不自然である．この存在単語の自然さを意味スコアとして算出し，誤り検出に用いる．しかし，意味スコアは，どの単語と共起しても不自然でない機能語に対しては意味をなさないため，本稿では内容語として名詞，動詞，形容詞のみに意味スコアを与える．内容語 w の意味スコア $SS(w)$ の算出手順は次の通りである．

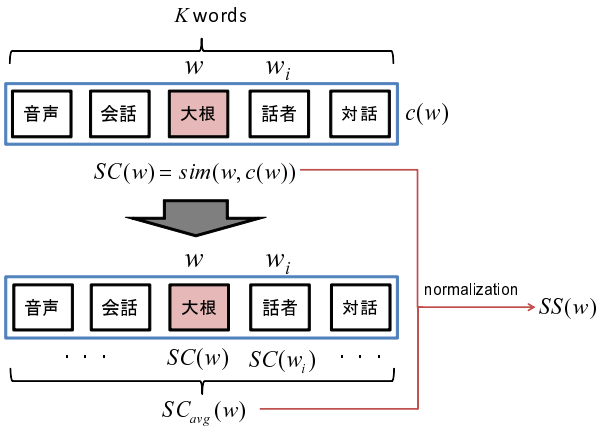

Fig. 1　意味スコアの算出

1. w の周辺に現れる内容語を，Fig. 1 のように文脈窓幅 K で集め，単語集合 $c(w)$ とする (w 自身も含む)．

2. 集合 $c(w)$ 内の各単語 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，$SC(w_i)$ とする．

$$SC(w_i) = sim(w_i, c(w)) \qquad (3)$$

3. $SC(w_i)$ から，平均 $SC_{avg}(w)$ を求める．

$$SC_{avg}(w) = \frac{1}{K} \sum_{i} SC(w_i) \qquad (4)$$

4. $SC(w)$ と $SC_{avg}(w)$ の差を意味スコア $SS(w)$ とする．

$$SS(w) = SC(w) - SC_{avg}(w) \qquad (5)$$

単語間類似度 $sim(w_i, c(w))$ の算出には，Latent Semantic Analisys (LSA) [4] を用いた．LSA は大量のテキストにおける単語の共起関係を統計的に解析することで，学習データに直接の共起がない単語間の類似度についても求めることができる手法である．

*Study on Features in Error Correction of Speech Recognition Result Based on CRF, by Ryohei Nakatani, 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	13
発話数	39,808	4,771
単語数	361,513	39,822

2.3　長距離言語情報と他の素性の組み合わせ

本稿では，長距離言語情報を他の素性と組み合わせることで，音声認識率の改善を狙う．単語の文脈一貫性と，音響尤度を組み合わせることで，認識精度の低い音声ドキュメントを棄却する研究 [5] が行われている．そこで，我々は，文脈一貫性の代わりに長距離言語情報を，音響尤度の代わりに Confusion Network 上の信頼度 (以降 CN 信頼度と呼ぶ) を用い，それらを組み合わせることで新たな素性とした．これらの素性を組み合わせることで，どれだけ文脈に合っていても，信頼度が低い場合は無視するなど，柔軟に特徴を選択することを目的としている．

3　評価実験

3.1　実験条件

音声認識器 Julius [6] の認識結果に対して，提案した素性を用いて誤り訂正の評価実験を行った．データは日本語話し言葉コーパス (CSJ) を用いた．音響モデル，言語モデルともに CSJ から学習した．LSA の学習には，CSJ の書き起こし文書のうち，評価データを含まない 2,672 講演を用いた．意味スコアを求める際の文脈窓幅は $K = 3$ とした．CRF による誤り検出モデルの学習と評価実験に用いたデータは，Table 1 のようになっている．学習する素性は，表層単語 bigram，trigram，CN 信頼度，LSA による意味スコア，そして 2.3 節で述べた素性を用いる．

3.2　実験結果

実験結果を Table 2 に示す．「SUB」は置換誤り，「DEL」は削除誤り，「INS」は挿入誤りの数をそれぞれ表している．「COR」は正解単語の数，「WER」は単語誤り率である．「CN-oracle」は，Confusion Set において常に正解の単語を選択したときの WER である．ただし，正解がないときはその Confusion Set 中で最も信頼度の高い単語を選んでいるため，ヌル遷移が選択されることで削除誤りが最小にはなっていない．「CN-best」は，誤り訂正前のベースとなる，Confusion Network の最尤候補列の WER で，「Nonsemantic」は，N-gram と CN 信頼度を素性とした際の，誤り訂正結果を表している．また，「Semantic」は「Nonsemantic」に意味スコアを追加した場合，「Proposal」は，意味スコアを単独で用いるのではなく，CN 信頼度と組み合わせて用いた場合の誤り訂正結果である．

Table 2　単語誤り率と誤り種類別の評価

	SUB	DEL	INS	COR	WER
CN-oracle	1,807	2,172	831	35,487	12.08
CN-best	7,198	1,834	3,423	30,434	31.28
Nonsemantic	6,416	2,405	2,213	30,645	27.71
Semantic	6,360	2,387	2,223	30,719	27,55
Proposal	6,285	2,441	2,149	30,740	27.31

表より，意味スコアを単独で用いた Semantic は，Nonsemantic と比較して WER が 0.16 ポイント改善している．一方で Proposal は，Semantic よりもさらに WER が 0.24 ポイント改善しており，意味スコアと CN 信頼度の組み合わせを素性とすることで，意味スコア単独よりも有効に働かくことが確認できた．また，CN-best と比較すると，トータルで 3.97 ポイント改善した．しかし，CN-oracle と比較すると，まだ改善の余地がある．

4　おわりに

本稿では，誤り訂正を行う際の素性として，長距離言語情報と認識信頼度を組み合わせて新たな素性とし，音声認識誤り訂正を行った．評価実験の結果，長距離言語情報を単独で用いるよりもさらに WER が改善した．

今後の課題として，誤り訂正の精度を向上させるため，N-gram や認識信頼度，長距離言語情報と競合しない，新しい素性を調査していく必要がある．

参考文献

[1] J. Lafferty, *et al.* “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” ICML, pp. 282-289, 2001.

[2] 中谷, 他, “文脈特徴を用いた CRF による音声認識誤り訂正”, 音講論 (秋), pp. 189-190, 2011.

[3] L. Mangu, *et al.* “Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks,” Computer Speech and Language, pp. 373-400, 2000.

[4] Jerome R. Bellegarda, “Latent semantic mapping,” IEEE Signal Processing, 5(22), pp. 70-80, 2005.

[5] 浅見, 他, “単語の文脈一貫性と音響尤度を用いた音声ドキュメント認識信頼度の推定”, 信学技法, SP2010-42, pp. 43-48, 2010.

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