N-best のスコアと文字列類似度を素性に用いた音声認識結果における誤り検出の検討

島田 渉平[1] 狩野 芳伸[1]

[1]静岡大学大学院総合科学技術研究科
{s-shimada, kano}@kanolab.net

1 はじめに

昨今 Google Home や Amazon Alexa, Siri のような音声認識を用いたアシスタントが一般的になりつつある. 近年の深層学習を用いた音声認識システムの発達により, このような音声認識を用いたアシスタントの実用性が十分高まっている. これらの音声認識アシスタントはユーザーの音声によって様々なコマンドを実行するものであるが, 声を用いて入力する以上どうしても音声認識誤りは発生する. しかし, これらの音声認識アシスタントは, まだまだ認識誤りに対する棄却が自然なレベルであるとは言いづらい. いわゆる, 我々人間が普段引き起こすような聞き間違いの機能が, これらのアシスタントにはまだ十分に備わっていないように感じられる.

本論文ではこのような音声認識システムにおいて, 自然な聞き間違いを実現するために, 音声認識された結果から, 我々人間が見た時に不自然と感じられる認識結果を判別し, 棄却するためのシステムについて提案する. 本論文における提案手法では, 認識結果の N-best リストを出力し, 各 N-best のスコア, 及びに各 N-best 候補間の文字列の類似度, ジャロ・ウィンクラー距離を学習素性とし, サポートベクターマシン(SVM)により, Word Error Rate(WER)の高い発話の検出を試みた. 実験の結果, 高い精度は得られず, より良い学習素性の検討が必要であると示された.

2 関連研究

音声認識精度を向上させるため, また対話誘導を行うために誤りと思われる認識結果を棄却するための先行研究はいくつか存在している. 駒谷ら[1]は N-best の出力文とスコアを用いて, 内容語に対する信頼度(Confidence Measure)を定義し, その値により対話戦略を変更するという研究を行なっている. 本研究では, スマートフォンなどにみられる誤入力補正のように, 後段でさらに認識結果を新たな認識結果へと変換できるよう, 内容語単位ではなく, 発話単位でスコアを算出する.

表 1 N-best 出力の例

ID	N-best	score
A01F0055_0001		
results-	上司に発達齢差が…	-6.07290055E+08
	調子に発達齢差が…	-6.07384705E+08
	朝市に発達齢差が…	-6.07290055E+08
	聴取に発達齢差が…	-6.07490064E+08
	長針に発達齢差が…	-6.07552017E+08
	長身に発達齢差が…	-6.07604500E+08
	銚子に発達齢差が…	-6.07632911E+08
	城氏に発達齢差が…	-6.07633091E+08
	浄水に発達齢差が…	-6.07670390E+08
	女子に発達齢差が…	-6.07823463E+08
results-	旋律のじょ	-1.68129270E+08
	県立のじょ	-1.68602883E+08
A01F0055_0002		
…		

3 提案手法

提案するシステムの概要を図 1 に示す. 本研究では音声認識デコーダーより出力された N-best 認識結果とその信頼度スコアより学習素性を抽出し, それを SVM 識別器へ出力, SVM によって正しい認識結果と判別されればそのまま出力, 誤った認識結果であると判別されれば棄却され, 後段の処理へと出力される.

本研究では SVM に入力する素性として, 出力された N-best 文の信頼度, また N-best 文それぞれの文字列の類似度を用いた.

3.1 N-best からの信頼度スコアの算出

音声認識デコーダーからは N-best を出力させる. N-best 出力の例を表 1 に示す. 本研究ではこの N-best のスコアの散らばりの度合いで, 音声認識の信頼度の算出を試みる. N-best は N=10 で出力を行なった.

音声認識デコーダーより出力された N-best 認識結果について, 最尤のスコアと他の N-best 候補との差を求め, それらのスコアの標準偏差を信頼

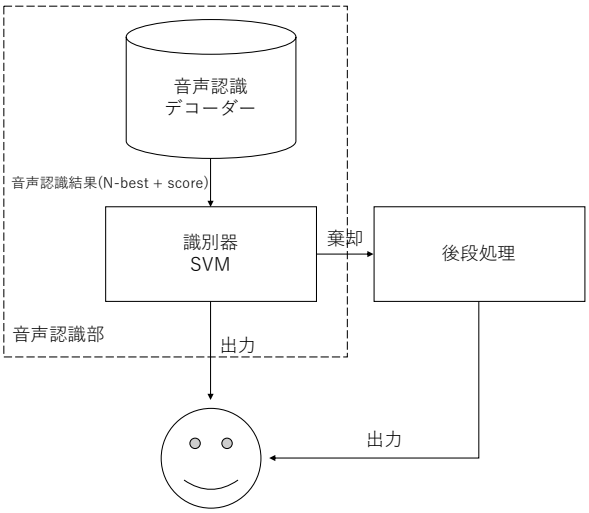

図 1 システムイメージ

度スコアとする．また，表 1 では一発話内の認識が二つに分かれている．このように，入力される一発話に対して出力が複数に渡る場合は，算出されたそれぞれの信頼度スコアの平均値を一発話の信頼度スコアとする．

3.2 ジャロ･ウィンクラー距離による文字列類似度の算出

SVM に入力するもう一つの素性として，最尤の N-best 文とそれぞれの N-best 文間の文字列類似度を用いた．本研究ではジャロ・ウィンクラー距離[2]を用いる．

出力される N-best について，最尤の N-best 文とそれぞれの N-best 文間のジャロ・ウィンクラー距離を算出し，算出されたジャロ・ウィンクラー距離の標準偏差を文字列類似度とする．また信頼度スコアと同様に，認識結果が複数に渡る場合は，各区間の文字列類似度の平均値を，一発話における文字列類似とする．

4 実験

4.1 SVM の学習

SVM の学習のために，CSJ より対話データを含まない 1260 講演分の音声を用いてあらかじめ音声認識を行なった．認識結果のうち，WER が 50%を超えたものには誤り認識のラベルを，それ以外には正しい認識のラベルを振り分けた．その中より誤り認識とした音声 3139 文と正しいと認識した 3139 文，合わせて 6081 文のデータを用いて学習を行なった．SVM のカーネルには Radial Basis Function (RBF)を用い，ハイパーパラメータはグリッドサーチにより，{C=1, gamma=0.1}に定めた．

表 2 実験結果

テストセット	Recall	Precision	F-Score
1	65.3	19.7	30.2
2	75	15.7	25.9
3	71.9	11.2	19.4

表 3 混同行列: テストセット 1

		Predict	
		誤り	正解
Actual	誤り	47	25
	正解	192	283

表 4 混同行列: テストセット 2

		Predict	
		誤り	正解
Actual	誤り	45	15
	正解	327	242

表 5 混同行列: テストセット 3

		Predict	
		誤り	正解
Actual	誤り	23	9
	正解	228	283

4.2 認識実験

提案手法により SVM を学習し，認識実験を行なった．データセットには日本語話し言葉コーパス(CSJ)[3]を用いた．

音声認識デコーダーには CMU Sphinx4[4]を用いている．音響モデル・言語モデルの構築には同じく日本語話し言葉コーパスを用いた．音響モデルは対話データを除いた 3209 講演分の音声データを用いて構築した．学会講演と模擬講演からなる 30 講演分のテストセットを用いたテストの結果，WER は 24%であった．言語モデルの構築には，テストセットと朗読音声を含まないデータを用いている．ここで朗読音声を含まなかった理由は，一部朗読音声中に学会講演の再朗読データが含まれていたためである．

認識実験に用いたテストセットは，音響モデル構築時と同様のものを用いた．30 講演分のデー

タを 3 つのテストセットに分け, それぞれで認識精度の実験を行なった.

5 実験結果

実験結果を表 2 に, それぞれのテストセットにおける混同行列を表 3〜5 に示す. 表 2 に示したように, F 値は想定していたよりも大幅に低い値となってしまった. 再現率の値はある程度の数値を保っているが, 適合率の値が全体的に低い数値となっている. それぞれのテストセットにおける混同行列からも見て取れるように, かなり多くの正解発話を誤り発話と分類してしまっている.

6 考察

再現率の値がある程度の数値を保っていることから, 学習した素性は事例に対して当てはまる部分もあるが, その他の事例に対しても当てはまっていることを意味している. この結果から, 誤り発話を検出するためには, これらの学習素性だけでは不十分, または不適格であると考えられる. 実際の学習データにおける誤り発話の数値を見てみても, N-best スコア, 文字列類似度のそれぞれの標準偏差がどちらとも高い数値を示しているもの, そのどちらもが低い数値を示しているもの, またどちらか片方のみが低い数値を示しているものなどが混在しており, これらの学習素性と実際の結果の相関性は低く感じられる.

また, 実験の際に用いたテストセットはどれも誤り発話の割合が低い, アンバランスなデータセットになっている. テストセットに用いるデータセットの分布も揃えて検討してみる必要性がある.

今後より良い結果を得るためには, まず学習素性を再び検討し直す必要があると考えている. また, LSTM などの再帰ニューラルネットを用いてモデルを構築するなど, N-best の出力値が音声認識の信頼度として有意な値を出力することが可能か, さらなる検証を進めていきたいと考えている.

7 おわりに

本稿では誤り率の高い音声認識結果を棄却するために, 音声認識デコーダーの出力する N-best のスコアと, 各 N-best 文の文字列類似度を素性に, SVM を用いた高 WER 発話の検出を試みたが, あまり高い精度は得られなかった.

今後統合を考えている後段の補正処理の精度によっては, WER の改善も考えられるが, より良い素性の選択は今後検討するべき事項であると考えている.

参考文献

[1] 駒谷和範, 河原達也, “音声認識結果の信頼度を用いた効率的な確認・誘導を行う対話管理”, 情報処理学会論文誌, Vol.43, No.10, pp.3078–3086 (2002).

[2] William E, Winkler., “String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, Proceedings of the Section on Survey Research Methods”, American Statistical Association, pp.354–359 (1990).

[3] 国立国語研究所, “国立国語研究所報告書 124 日本語話し言葉コーパスの構築法”, 2006.

[4] Paul Lamere, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred K. Warmuth, Peter F. Wolf, “The Cmu Sphinx-4 Speech Recognition System”, 2001