6N-03

プレゼンテーションにおけるスライド情報を用いた音声認識結果の自動修正

神谷 賢太郎† 川瀬 卓也† 東中 竜一郎† 長尾 確†

名古屋大学 大学院情報学研究科†

1. はじめに

近年，音声認識技術の向上により音声認識を用いた議事録作成が可能になりつつある．しかし，会議におけるプレゼンテーションなどでは，認識が難しい専門用語が多く使用されることなどから正確な議事録作成は困難である．

音声認識を用いた自動議事録作成を目的に，我々はプレゼンテーションでの発表者と参加者の音声を収集している．クラウド音声認識を利用して収集した音声を認識させたところ，Word Error Rate (WER) が 26.4%と必ずしも良いとは言えない結果となった．発表者や参加者の発言に専門用語が多く含まれることや，音声に発言以外の環境音が混在していることなどが精度低下の原因である．プレゼンテーションではスライドを使用する機会が多く，また複数の発言者が同じ議題について議論を行っている．このようなプレゼンテーションの特徴から，スライド情報や認識対象の周辺の発言（以下，周辺発言情報）が音声認識精度の改善に活用できると考えられる．

そこで本研究では，スライド情報や周辺発言情報を用いた認識結果の修正手法を提案し，音声認識精度の改善を試みる．

2. 提案手法

本システムの構成を図 1 に示す．学習フェーズでは，スライド情報や周辺発言情報から補足情報を抽出し，認識結果，補足情報，外部データを用いてモデルの学習を行う．評価フェーズでは，学習フェーズ同様に抽出した補足情報をもとに学習済みモデルを用いて認識結果の修正を行う．

2.1. 補足情報の抽出とフィルタリング

スライド情報や周辺発言情報から補足情報の抽出を行う．補足情報の対象は専門用語を多く含む名詞や動詞などの内容語とする．抽出した補足情報にはスライド情報や周辺発言情報に含まれる全ての内容語が含まれており，その中にはノイズとなり得る単語も含まれている．そこで，補足情報のフィルタリングを行う．

除去すべき単語を把握するため，認識誤り単語の分析を行った．その結果，多くの認識誤り単語と正解単語には，「最初」と「最小」のような発音的類似関係があることが分かった．会議の音声の認識結果と，それに付随する情報から発音的類似関係に着目して情報抽出を行う研究がある [3]．この研究における情報抽出の手法を参考に，認識結果と補足情報に含まれる各内容語の発音的類似関係に着目して補足情報のフィルタリングを行う．

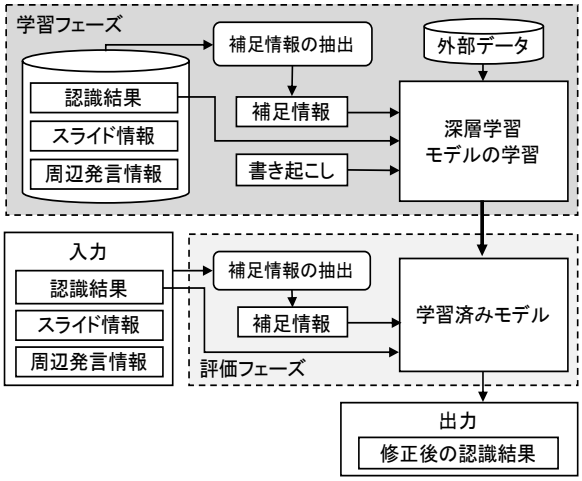

図 1 認識結果修正システム

発音的類似関係の有無の確認には IPA 距離 [1] を利用する．IPA 距離は，単語を発音記号列に IPA 発音記号を用いて変換し，発音記号列間の編集距離から計算する．算出した IPA 距離がある一定の閾値以上となる単語を除去する．

2.2. 深層学習モデル

認識結果の修正を行う深層学習モデルの実装には，機械翻訳結果を事後編集するタスクで利用される OpenNMT-APE [2] を用いた．ここで用いられるモデルは機械翻訳結果を入力とし，事後修正した翻訳結果を出力する．このモデルを，認識結果を修正するタスクに適応させるために，認識結果と補足情報を入力とし，修正した認識結果を出力するようにファインチューニングを行う．

3. 評価実験

補足情報をもとに認識結果を修正する手法の有効性を実験により検証した．

3.1. データセット

本研究室で実施されたプレゼンテーションでの発言から音声データを収集し，書き起こしを行った．また，Microsoft が提供している Azure Speech to Text (以下，Azure) を利用して，収集した音声データの認識結果を取得し，書き起こしと合わせて認識結果修正データセットを作成した．このデータセットを

Automatic Correction of Speech Recognition Results in Presentation by Using Slide Information
† KAMIYA, Kentaro (kamiya.kentaro@b.mbox.nagoya-u.ac.jp)
† KAWASE, Takuya (kawase@nagao.nagoya-u.ac.jp)
† HIGASHINAKA, Ryuichiro (higashinaka@i.nagoya-u.ac.jp)
† NAGAO, Katashi (nagao@nuie.nagoya-u.ac.jp)
† Graduate School of Informatics, Nagoya University

15:1:1 で分割し，15 を学習，1 を検証，残りの 1 を評価に用いた．

また，本研究室で提案・運用しているディスカッションマイニング [4] という会議記録システムで蓄積されたテキストデータ(以下，DM データ)がある．DM データはプレゼンテーションでの議論内容を要約したテキストデータであり，データ量は約 25,000 文である． DM データは発言の音声に対する書き起こしではなく，議論内容を要約したものなので，モデルのファインチューニングには直接使用できない．しかし，今回は生成モデルを学習しており，言語モデルを特定のドメインのデータに適応させておく方が，適切な生成ができると考えた．そこで，DM データを用い，事前学習として，議論内容の要約のオートエンコーダを学習するという過程を挟むことで，言語モデルを研究内容に適応させることが可能だと考えられる．

3.2. 比較手法

認識結果修正データセットを用いて，以下(a)-(e)の 5 パターンの手法を比較した．提案手法は(c)-(e)である．

(a) クラウド音声認識: Azure を用いて音声認識を行う．ここで取得する認識結果は(b)-(e)の入力データとして使用する．

(b) 単純な事後編集による手法: (a)の認識結果をモデルに入力し，修正後の認識結果を出力する．

(c) スライド情報を使用した事後編集による手法: (a)の認識結果に加え，スライド情報から抽出した補足情報をモデルに入力し，修正後の認識結果を出力する．入力である認識結果と補足情報は，[SEP]というタグ区切りで連結する．

(d) 周辺発言情報を使用した事後編集による手法: (c)と基本的には同じだが，補足情報はスライド情報ではなく周辺発言情報から抽出する．

(e) スライド・周辺発言情報を使用した事後編集による手法: (c), (d)と基本的には同じだが，補足情報はスライド情報と周辺発言情報の両方から抽出する．

上記 5 パターンの手法に加え，(c)-(e)については，(1) IPA 距離を用いた補足情報のフィルタリングを行うかどうか，(2) DM データを用いた事前学習を行うかどうか，の各組み合わせについても同様に比較した．また，IPA 距離の閾値は 0.25 に設定した．

3.3. 評価結果と考察

提案手法の評価結果を表 1 に示す．(e)のフィルタリングなしで，事前学習ありの場合に WER が最も良い結果であり，(a)の修正前であるクラウド音声認識と比べ約 2.9 ポイントの改善がみられた．ここで，フィルタリングがある場合の方が，ない場合に比べて修正後の認識結果の精度が低いことから，フィルタリングにより，重要な特徴を持つ補足情報を誤って除去している可能性が考えられる． (c)と(d)を比較すると，(c)の方が全てのパターンにおいて修正後の認識結果の精度が良い．この結果から，周辺発言情報よりもスライド情報の方が認識結果の修正に重要となる情報をより多く含むと考えられる．

表 1 提案手法の評価結果
フィルタリング:補足情報のフィルタリングの有無
事前学習:DM データを用いた事前学習の有無

	フィルタリング, 事前学習	WER (%)
(a)クラウド音声認識	-	25.71
(b)単純な事後編集	-	23.95
(c)スライド情報を使用した事後編集	なし，なし	23.59
	なし，あり	22.87
	あり，なし	23.73
	あり，あり	22.90
(d)周辺発言情報を使用した事後編集	なし，なし	24.69
	なし，あり	23.19
	あり，なし	23.79
	あり，あり	22.95
(e)スライド・周辺発言情報を使用した事後編集	なし，なし	23.94
	なし，あり	**22.81**
	あり，なし	23.43
	あり，あり	22.86

補足情報を使用した(c)-(e)の各 4 パターンのうち，WER が最も良いパターンと，補足情報を使用していない(b)との間に差があるかどうかをウィルコクソンの符号順位検定で確認したところ，(c)-(e)の全てにおいて有意差があることを確認した．また，スライド・周辺発言情報を使用した(e)の場合の有意差が最も顕著であった．

4. まとめ

本研究では，スライド情報や周辺発言情報から補足情報を抽出し，補足情報をもとに認識結果を自動修正する手法を提案した．認識結果の修正には深層学習モデルを用い，認識結果と補足情報を入力とし，修正した認識結果を出力するようにニューラルネットワークのモデルを学習させた．実際のプレゼンテーションデータを用いた評価実験の結果から，修正後の認識結果の精度が修正前と比べ約 2.9 ポイント改善できた．今後は，より効果的な補足情報の抽出アルゴリズムを検討する予定である．

参考文献

[1] 河野宏志，城塚音也，高木徹．国際音声記号を用いた発音類似度算出アルゴリズムの検討．情報科学技術フォーラム講演論文集 13.2,pp.261-262, 2014.

[2] Gonçalo M, et al., A simple and effective approach to automatic post-editing with transfer learning. In Proc. ACL, pp.3050-3056, 2019.

[3] 六浦由香，大平茂輝，長尾確．クラウド音声認識と会議リソースに基づく半自動議事録作成．情報処理学会第 82 回全国大会講演論文集，pp. 47-48, 2020.

[4] K. Nagao et al., Discussion Mining: Annotation-Based Knowledge Discovery from Real World Activities. In Proc. of the Fifth Pacific Rim Conf. on Multimedia, pp.522-531, 2004.