様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月　 ８ 日現在

機関番号：１４３０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ１１３５４

研究課題名（和文）ニューラルネットワーク言語モデルの適応的な自動構成法

研究課題名（英文）Automatic adaptation framework of neural network language model

研究代表者
秋田　祐哉（AKITA, Yuya）

京都大学・経済学研究科・教授

研究者番号：９０４０２７４２

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：音声認識において，一般的な話題のテキストデータから学習したモデルでは，講義・講演のような専門性のある内容の音声を高い精度で書き起こすことは難しい．このための方策として，言語モデルを特定の話題に適応する，いわゆる言語モデル適応がある．本研究では，認識対象の音声とともに与えられる話題関連のテキストを用いて，ニューラルネットワークに基づく言語モデルに対して自動的に適応処理を行い，自動的に音声の字幕を作成するシステムを構成した．本システムは，適応したモデルによる事後的な字幕の作成だけでなく，リアルタイムの字幕付与も行うことができる．

研究成果の学術的意義や社会的意義
音声認識はコミュニケーションの支援技術として社会的な重要性が増大しているが，専門的な内容を含む音声に対してニューラルネットワークのような高度なモデルを適用することには技術的な困難がある．本研究により，非専門家がより性能の高い音声認識を容易に取り扱えるようになることには，大きな意義があると考えられる．

研究成果の概要（英文）：In automatic speech recognition, common models trained with general texts have limited performance for specialized topics, such as those in classroom lectures and academic talks. To deal with this problem, language model adaptation is often conducted. In this study, we investigated automatic adaptation framework of neural-network-based language models by using texts relevant to the topics in the target speech, and incorporate it into our system of automatic captioning, which produces captions for both of recorded audio and real-time audio, with the adapted language models.

研究分野：音声認識

キーワード：音声認識　ニューラルネットワーク　言語モデル

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年の音声認識は，人のコミュニケーションを支援するツールとしての役割を拡大している．たとえば講義・講演などの映像・音声配信において，音声の聴取が難しい利用者を支援するために，あるいは内容の理解を促進するために，書き起こしを作成したり字幕を付与したりする取り組みがある．また，このような事後的な字幕付与だけでなく，リアルタイムの字幕の提供を行う取り組みも進められている．音声認識の導入により，従来必要としていた人手・労力が削減され，より多くの場面でコミュニケーション支援を行えることが期待できる．

音声認識は，音声の音響的特徴（スペクトル等）と言語的特徴（語彙・構文等）をそれぞれ音響モデル・言語モデルとして統計的にモデル化して実現されてきた．したがって，音声認識が可能なのは，これらによりモデル化された範囲に限定されている．すなわち，日常会話を想定した，あるいは専門的話題を前提としない「浅く広い」モデルでは，講義・講演のような専門性の高い音声を十分に書き起こすことは不可能である．これは，音響的な不一致もあるが，言語モデルが話題をカバーできていないことが致命的な問題として考えられる．

このための方策として，言語モデルを特定の話題に適応する，いわゆる言語モデル適応の研究が行われてきている．これらの手法は，比較的大量の学習データを前提としたり，あるいはパラメータの調整を必要としたりすることが一般的である．これに対して，前述した実際の利用場面では，たとえば当該講演の予稿やスライドのみが利用可能で，調整に利用可能なテスト用音声もなく，しかも専門家による十分なチューニングの余裕がない．つまり対象ごとに音声認識システムをカスタマイズすることは容易ではなく，非専門家に利用してもらう際にこの点が支障となっている．

２．研究の目的

本研究では，講義・講演のような専門的内容に対する音声認識を容易に実行可能とするため，音声認識システムを自動的に構成する枠組みについて取り組む．講義・講演の書き起こしや字幕作成のために音声認識を行う際，これらに出現する専門的話題をカバーした音声認識システムをあらかじめ構成(適応)する必要があるが，通常このために利用できるデータは限られており，また専門家でなければ作業は容易ではない．そこで，モデルの形態や性能指標，また少量のデータでも有効な適応手法を検討し，自動的な適応・構成法を実現する．音声認識を用いた字幕作成タスクにこれらの枠組みを導入し，専門家ではない一般の利用者による字幕作成を通じて有効性を示す．

３．研究の方法

本研究は，我々が開発してきた音声認識に基づく自動字幕付与システムをプラットフォームとして実施した．本システムでは，ユーザにより収録された講義・講演や討論などの音声・映像に対して，事後的に字幕を付与することを想定している．まず，ユーザがこれらのコンテンツを字幕サーバにアップロードする．音声・映像に加えて，言語モデルを話題に適応させるために，コンテンツの話題と関連するテキスト（たとえば講演予稿やスライド）もアップロードすることができる．字幕サーバではコンテンツからの音声の抽出および検査が行われ，ユーザの指定や関連テキストに応じて自動的に音声認識システムが構成された上で認識処理が実行される．

本研究ではこのシステムを拡張して，講義・講演の会場で情報保障のためにリアルタイムに字幕を作成・表示するシステムも構築した．本システムは，講義・講演会場で作業者が入出力・編集に使用する PC と，音声認識を行うサーバから構成される．講師の音声は PC に入力され，PC 側で発話検出・セグメンテーションを行ったのち，サーバにネットワーク経由で送信される．サーバではあらかじめ対象の講義・講演用に構成された音響モデル・言語モデルを使用して音声認識を行い，この結果をネットワーク経由で作業者の PC に送信する．サーバでは GPU を用いて計算することにより，実時間以下の処理時間で音声認識を行っている．

本研究では言語制約としてニューラルネットワーク言語モデルを導入した．モデル化の単位としては単語を想定している．ニューラルネットワーク言語モデルでは，入力されるベクトルは語彙のサイズだけの次元数がある．通常はこの入力層の後に次元を圧縮する射影層(Embedding)をおき，続いて隠れ層としてリカレント構造を持つユニットをおいて，過去の入力（履歴）に関する情報を保持する．本研究の実際のユニットとしては LSTM が用いられる．出力層は語彙のサイズだけの出力を持ち，語彙中の各単語の確率を出力する．この枠組みから明らかなように，この確率は入力と履歴の状態により変動する．隠れ層が保持している履歴は，原理的には過去のすべての入力についての情報を保持していることから，ニューラルネットワーク言語モデルは統計的（N-gram）言語モデルよりも遠くの過去の情報を用いて単語を予測することが可能である．一般的に文の生成では，直近に用いられた単語だけでなく，以前に出現した単語をもとに次に用

いられる単語が決定されることもあるので，長い履歴を保持できることによる予測性能（音声認識性能）の改善が期待される．

ただし，入力と出力の次元数（語彙のサイズ）は学習の段階で固定される．このため，いったん学習したモデルについて，たとえば単語を追加するなどのために次元数を変更することは容易ではなく，再学習が必要となる．また，ニューラルネットワークに一般的な傾向として過学習があり，言語モデルでもニューラルネットワークのモデルのみを用いることはかえって性能が低下することが少なくない．さらに，ニューラルネットワーク言語モデルは N-gram モデルと比較して計算コストが大きく，多数の文候補（仮説）を生成しつつ評価する認識処理に単純に適用することは処理の大きな遅延の要因となりうる．そこで本システムでは，ニューラルネットワーク言語モデルと N-gram モデルとの補間を行って確率を算出する．本研究では，適応のために与えられたテキストを，ベースとなる（適応前の）モデルの学習テキストと組み合わせてニューラルネットワーク言語モデルの学習を行うともに，N-gram モデルの確率も更新する．すなわち，同じテキストから，適応されたニューラルネットワーク言語モデルと N-gram モデルがそれぞれ構築される．

音声認識の際は，ニューラルネットワーク言語モデルの計算が高コストであることから，まず N-gram 言語モデルのみを言語制約としていて認識結果をいったん生成する．認識処理の際に検討された文の候補（仮説）には，それぞれ音響モデル・言語モデルに基づくスコアが与えられており，このスコアによって最善の仮説が認識結果として出力されるが，ここでは 2 番目以降のものも多数保持しておく（いわゆる N-best 文）．次に，これらの認識結果の各文についてリスコアリングを行う．すなわち，各文のスコアのうち言語モデルによるスコアについてニューラルネットワーク言語モデルを反映させて計算し直し，得られたスコアにしたがって認識結果の N-best 文を並べ直して，最もスコアの高い文を最終的な認識結果として得る．

本システムでは，これまでに述べたような従来の音声認識で個別に構築されていたモデルを 1 つのニューラルネットワークに統合して学習・認識する，いわゆる End-to-End 型の音声認識も搭載した．従来の音声認識では，音響モデル・言語モデルおよび単語辞書をもとにデコーダが入力音声に対して多数の仮説を展開して探索するため必然的に時間がかかり，リアルタイムに認識するためには種々の高速化の工夫，また探索空間の制限のような性能低下につながる対応が必要である．これに対して End-to-End 型の音声認識では，入力音声をディープニューラルネットワークに投入して，ネットワークの計算（推論）を行うだけで認識結果が出力されるため，実行の制御はシンプルであり，高速に認識を実行できる．

４．研究成果

本システムで導入した枠組みについて，実際の字幕付与データにより性能評価を行った．用いたデータは京都大学で行われたシンポジウムの講演で，音声認識に基づく字幕付与が行われた．この講演の予稿を使用して言語モデルを適応し，あらためて音声認識を行ってその結果を字幕のテキストと比較することで認識性能の評価とした．

本実験では，モデルが複雑になると学習時間が大きく増加することから，モデルができるだけシンプルになるよう，あらかじめパラメータ（リカレント層の種類，隠れ層の数，射影層のユニット数）の比較検討を行って定めた．なお，適応テキストとベースの学習テキスト（CSJ 学会講演データ）を結合した後の語彙サイズ（つまり入力・出力ベクトルの次元数）は 36,699，データの総単語数は 798 万単語である．モデルの学習に要した時間は 39 分である．学習エポック数は事前にベースモデルの学習を通じて調整した．音声認識の結果，従来の統計的言語モデルに対して，ニューラルネットワーク言語モデルの導入により認識誤りが 3.3%削減され，導入の効果を確かめることができた．

End-to-End 型音声認識についても，実際の講演音声を用いて認識を行った．使用したデータは京都大学で開催されたシンポジウムで，講演者 2 名・計 62.6 分（実発話時間 47.9 分）である．従来の音声認識では認識に 12.5 分を要したのに対して，End-to-End 型音声認識では 4.5 分であった．音声は一定以上の無音で区切って認識したため，合計で 768 区間が入力され，平均の長さは 3.7 秒であるが，End-to-End 型ではこれらを平均 0.35 秒で認識しており，高速に認識できているといえる．

音声認識を提供するサービスは多くあるが，本研究のようにモデルをカスタマイズして非専門家の一般向けに提供する枠組みは本研究の実施時点では国内外で希有のものであり，実際の字幕付与の場面でも利用された．本研究の実施期間中，音声認識や自然言語処理の技術と性能は飛躍的に向上し，特に End-to-End モデルや大規模言語モデルの発展が著しい．今後はこれらに基づくさらなる方法論を探索することが必要であろう．

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 秋田祐哉・上乃聖・三村正人・河原達也
２．発表標題 自動字幕作成システムにおけるモデルの拡張
３．学会等名 情報処理学会アクセシビリティ研究会
４．発表年 2020年

１．発表者名 秋田祐哉・上乃聖・三村正人・河原達也
２．発表標題 音声認識を用いた字幕作成システムの改良
３．学会等名 情報処理学会アクセシビリティ研究会
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関