情報処理学会論文誌 Vol.61 No.3 608–617 (Mar. 2020)

音声認識エンジンに独立な特定ドメイン向け音声認識誤り識別器

田中 拓[1] 野田 直希[1] 加藤 恒夫[1] 山本 誠一[1]

受付日 2019年5月28日, 採録日 2019年11月29日

概要：大語彙音声認識の精度向上にともない，機械翻訳エンジンや対話エンジンと組み合わせた特定環境向けの音声翻訳システムや音声対話システムの実用化が検討されている．しかしながら，汎用的な音声認識エンジンは特定環境のドメインや文脈にそぐわない認識結果を出力する場合がある．このような場合に，特定ドメインや文脈の情報を用いて認識結果を検証し，誤認識の可能性が高い場合にはユーザへの出力を止める機能が求められる．従来，認識結果の検証機能は音声認識エンジンの内部変数を用いて汎用的な系列ラベリングを行うものであったが，本研究では，特定ドメインで得られた少量のログを用いて，音声認識エンジンとは独立に，発声ごとに認識結果に誤りが含まれるか否かの 2 値分類を行う音声認識誤り識別器を提案する．具体的には，タクシー車内で運用された音声翻訳システムの約 1 万発声分のログを用い，認識結果の単語系列を Bag-of-Words（BoW）で表現する識別器をベースラインとして，同単語系列を単語クラスを用いて Bag-of-Word-Classes（BoWC）で表現する識別器，注意機構付きリカレントニューラルネットワーク（RNN）に基づく識別器の 2 種類を検討した．BoWC で表現する識別器においては，シソーラスに基づく方法と単語分散表現に基づく方法を比較評価し，後者の方が優れることを確認した．RNN に基づく識別機では学習データの正例と負例の調節により精度改善を図った．評価の結果，正例と負例を調節した注意機構付き RNN に基づく識別器において，最も高い音声認識誤り識別精度 75.8%を得た．また，誤認識単語に対する注意機構の注視精度を評価した結果，80%を超える注視精度を確認した．

キーワード：音声翻訳システム，音声認識誤り識別，リカレントニューラルネットワーク

Speech Recognition Error Identifier in Specific Domain Independent of Automatic Speech Recognition Engine

Taku Tanaka[1] Naoki Noda[1] Tsuneo Kato[1] Seiichi Yamamoto[1]

Received: May 28, 2019, Accepted: November 29, 2019

Abstract: In accordance with the recent conspicuous development of automatic speech recognition (ASR) technologies, speech translation and spoken dialogue systems combining a generic ASR engine and a machine translation engine or a dialogue management engine are considered in various places. However, generic ASR engines sometimes output erroneous results which do not fit the domain or context. A preventing function of outputting such erroneous responses based on verification of ASR results is demanded by the systems. Although the verification of ASR results has been defined as a general sequence labeling task using internal variables of an ASR engine conventionally, we propose ASR error identifiers which estimate if a result contains errors or not independent of the ASR engine, trained with limited amount of log data of the domain. Specifically, we developed ASR error identifiers based on a random forest with Bag-of-Word-Classes (BoWC) features and another ASR error identifier based on a recurrent neural network (RNN) with an attention mechanism. We evaluated the performance of the ASR error identifiers comparing with a baseline method based on Bag-of-Words (BoW) features using log data of about 10 thousands utterances collected through a trial of a speech translation system operated in taxi. Two methods of composing BoWC were compared and one based on distributed representation of words outperformed the other based on thesaurus. The accuracy of RNN-based error identifier was improved by balancing positive and negative samples in the training data. Experimental results showed that the RNN showed the best accuracy of 75.8% with the balanced training data. Furthermore, the attention mechanism showed an accuracy of focusing on erroneous words over 80%.

Keywords: speech translation system, speech recognition error identification, recurrent neural network

1. はじめに

近年，大語彙音声認識の性能が飛躍的に向上したことで，汎用的な音声認識エンジンと検索エンジン，対話エンジン，機械翻訳エンジンを組み合わせ，音声検索システム，音声対話システム，音声翻訳システムなどの実用化が進みつつある．たとえば，言葉の壁を超えるコミュニケーションを支援する音声翻訳システムは外国人旅行者が常時携帯できるスマートフォンアプリ [1] のほか，駅などの交通機関，医療機関，店舗など特定のロケーションへの整備も期待されている．しかしながら，音声認識を妨げる要因は数多く存在し，たとえば，背景雑音の影響により認識精度は低下する．また，音声認識用の言語モデルは大量の学習データを必要とするため，たとえば国内観光のようなある程度の広さのドメインをカバーするように学習されるが，実際の利用環境ではドメインや文脈が限定されるため，特定のドメインや文脈にそぐわない認識結果が出力されることがある．音声翻訳システムであれば，誤認識により意思疎通が阻害されるため，音声認識が完了した時点で誤認識を検出し，再発声を促すなどの対策が必要である．

従来，誤認識検出の問題は音声認識結果に対する汎用的な検証問題として設定され，素性に単語系列とその事後確率に音声認識エンジンの内部変数も加えて系列ラベリング問題として解くものが多い．しかし，汎用的な音声認識エンジンを機械翻訳エンジンなどと組み合わせて特定環境向けのシステムを構成する場合，特定環境のドメインや文脈に沿った出力であるかを検証する機能が必要となる．ドメインに特化したそのような機能は，機能の汎用性と専用性の分離の観点から個別のエンジンには独立であって，内部変数を用いない方が望ましい．

本研究では，音声認識エンジンをブラックボックスと見なし，音声認識の単語系列のみを素性とする音声認識誤り検出手法を検討する．特定ドメインのデータとして，タクシー車内で試験運用された音声翻訳システムのログに含まれる発声の書き起こしと認識結果を利用し，認識結果がタクシードメイン以外の語彙を含むか否かの観点で，誤認識の有無を識別する音声認識誤り識別器の構成手法を検討する．具体的には，Bag-of-Words を素性とするベースライン手法に対して，単語クラス（BoWC）を素性とするランダムフォレストによる識別器，注意機構付き双方向 LSTM（Long Short-Term Memory）による識別器を設計，評価する．

ランダムフォレストの識別器においては語彙を表現する素性以外に表層的素性などの追加により改善を図るとともに，単語クラスの構成方法としてシソーラスに基づく方法と単語分散表現に基づく方法を提案し，比較評価を行った．注意機構付き双方向 LSTM においては学習データの正例と負例の割合を調整することで負例の検出精度を改善した．

注意機構は入力単語系列のどこに焦点を当てるべきかを教師なしで学習する．ログデータを分析した結果，誤認識は似た音を持つが意味的には遠い単語もしくは単語系列への置換として発生することが多いため，注意機構は誤認識の箇所に焦点が当たるように学習されると考えられる．そこで，注意機構の誤認識単語への注視精度を評価した．具体的には，認識結果の単語系列について注視の有無と正誤を比較し，注視による誤認識単語の検出精度を測った．

2. 関連研究

音声認識誤りの検出は，単語系列の各単語に対して正解か不正解かの 2 値を識別する系列ラベリングの問題として定義され，様々な素性と識別器が検討されてきた．

基本的な手法として，単語ラティスや N-best リストから求められる各単語の事後確率に対する閾値処理により 2 値判定を行う手法 [2], [3]，誤認識に対する帰無仮説の検定により 2 値判定を行う手法 [4], [5] がある．

さらに，事後確率以外の様々な素性を組み合わせて識別器を学習する手法がある．Tam らは，各単語の事後確率，N-gram 確率，他の単語候補の数，前後単語の事後確率などの基本的な素性に，RNN 言語モデル（Recurrent Neural Network Language Model，RNNLM）によるリスコアリングで計算される条件付確率や隠れ状態の変化量と，DNN 音響モデルによる認識結果と GMM 音響モデルによる認識結果の一致度を示す複数の素性を追加して，フィードフォワードニューラルネットワーク（Feed Forward Neural Network，FFNN）を学習し，2 値識別の精度を改善している [6]．Ogawa らは，音声認識エンジン内部の単語コンフュージョンネットワーク（Word Confusion Network，WCN）を単純化して，正解，置換誤り，挿入誤り，削除誤りの 4 種類のパスに集約する単語アライメントネットワーク（Word Alignment Network，WAN）を生成し，単語アライメントに関する情報を追加素性とした [7]．

識別器としては，単語ごとに判定を行う SVM や FFNN のほか，系列を考慮できる識別器として条件付確率場（Conditional Random Field，CRF）や RNN も用いられている．Ogawa らは CRF を用いた [7]．Korenevsky らは，WAN から取得した 4 種類の事後確率を追加素性として CRF と RNN を比較し，同等の識別性能を確認している [8]．さらに，Ogawa らは，音声認識結果の単語系列に加え，その品詞，事後確率，N-gram 確率などを素性として，多層の双方向 RNN（Deep Bidirectional RNN）を学習し，CRF や他の NN を凌ぐ識別精度を確認している [9]．

一方で，汎用的な音声認識エンジンを機械翻訳エンジンなどと組み合わせて特定ドメイン向けのシステムを構成する場合，個別のエンジンの内部変数は参照できないこ

[1] 同志社大学
Doshisha University, Kyotanabe, Kyoto 610–0394, Japan

表 1 運転手発話の統計量
Table 1 Statistics of drivers' utterances.

翻訳言語の種類	セッション数	発話数	単語正解精度	文（発話）正解精度	発話平均単語数	語彙サイズ
日英，日韓，日中	270	10,019	93.8%	59.0%	6.9	5,099

表 2 発話内容の分類と相対度数分布
Table 2 Classification of utterance contents and relative frequencies.

会話分類	説明	例	割合
基本案内	観光地など基本的な案内	正面の建物が県庁です	25.8%
応用案内	基本案内より詳しい説明	ここはホタルも飛んでいます	17.0%
その他	その他	日本は何回目ですか	15.7%
アナウンス	運転に関する内容	シートベルトのご協力お願いします	14.4%
あいさつ	挨拶など	こんにちは	4.7%
時間	時間に関する内容	二十五分ぐらいです	4.6%
食べ物	食べ物に関する内容	イカ	4.3%
機械翻訳	機械翻訳機に関する内容	後ろについている機械に話して下さい	3.5%
天気	天気に関する内容	台風が近づいております	3.1%
感想	感想	楽しんでいただけましたか	1.7%
注意	観光地などでの諸注意	展示しているものは写真を撮ったらいけません	1.5%
イベント	イベントの案内や説明	あなたに絵はがきのプレゼントがあります	1.4%
約束	出発時間などの約束	十一時までここで待ってます	1.1%
電車	電車や時刻に関する内容	汽車は何時ですか	0.6%
おみやげ	お土産に関する内容	おみやげですか	0.3%
宿泊	宿泊に関する内容	今日は近くに泊まるんですか	0.2%

とが多い．そして，認識結果が正しいか否かは，音声認識エンジンとは独立な情報を用いて検証する方が適切である．Errattahi らは，音声認識エンジンをブラックボックスととらえ，認識結果を構成する単語の信頼度（Confidence Measure），双方向の大語彙汎用 N-gram 言語モデルによる言語確率，同言語モデルから得られる文奇異性（Sentence Oddity）を SVM の入力として，単語レベルの誤認識の識別と分類を行う手法を提案している [10]．Errattahi らの手法は特定ドメイン向けではなく汎用を目指しており，特徴量も大語彙汎用 N-gram に基づき，最終的に音声認識エンジンが出力する単語事後確率も用いて単語系列ラベリングを行うため，本研究とは方向性が異なるが，単語レベルの誤認識の分類精度で 82%を得ている．

3. タクシー車内音声翻訳システムの対話ログ

特定ドメインのデータの一例として，タクシー車内で試験運用された言語音声翻訳システムのログデータを用いた．

3.1 タクシー車内多言語音声翻訳システム

多言語音声翻訳システムは，地方都市のタクシーにおいて観光スポットを周遊する外国人の乗客が希望した場合に使用された．対応言語は日本語，韓国語，中国語，英語の 4 種類である．

運転手と乗客は運転席と後部座席にそれぞれ設置されたタブレット端末とスピーカを用いて対話を行う．運転手はハンドルに装着された操作スイッチを押して音声を入力する．運転席のタブレット端末には音声認識結果，翻訳結果，発話者の言語への再翻訳結果が表示される．後部座席のタブレットには翻訳結果が表示され，スピーカから音声が出力される．乗客は，タブレット端末のアプリケーションのボタンを押して音声を入力する．音声認識結果に対する翻訳結果はタブレットにテキスト表示され，スピーカから音声が出力される．

3.2 音声翻訳システムの対話ログデータ

試験サービスにおいて記録された多言語音声翻訳システムの対話ログの発声内容を分析した．複数名の運転手と数百名の外国人の乗客の間で実際に交わされた自発的な対話，270 セッションである．対話ログは運転手と乗客の対話に沿って，人手による発話の書き起こし，音声認識結果，機械翻訳結果の組で構成される．運転手の発話が乗客の発話よりも多く，運転手発話に対する音声認識や機械翻訳の誤りにより乗客に意図が伝わらない状況が比較的多かった．

複数名の運転手による発話の基本的な統計量を**表 1** に示す．270 のセッションにより 10,019 発話が記録されたので，1 セッションあたりの平均発話数は約 37 である．語彙サイズは全運転手発話に含まれるユニークな単語の種類の数である．また，運転手発話の内容に基づく 16 種類の分類の相対度数分布を**表 2** に示す．観光に関する基本案内や応用案内などの発話が多く，地名やランドマークなどの固

表 3 音声認識誤りの分類と相対度数分布
Table 3 Classification of speech recognition errors and relative frequencies.

音声認識誤り種別	書き起こし文	音声認識結果	割合
a) 固有名詞の誤認識	これから イオン 北に向かいます.	これから 家を 北に向かいます.	8.1%
b) 同音異義語への誤認識	汽車 で 2 時間ぐらいかかります.	貴社 で 2 時間ぐらいかかります.	7.9%
c) 似た音素列の単語への誤認識	天候がよくなかったら 中止 にします.	天候がよくなかったら 上司 にします.	47.1%
d) 発話の一部分の脱落	タクシー の中は暑くありませんか.	____ の中は暑くありませんか.	17.9%
e) フィラーによる誤認識	えー，カニとか白イカ，マグロなど.	ええ，カニとか白いかマグロ.	1.0%
f) その他の音声認識誤り	駅 に行きます.	きゅう に行きます.	18.0%

表 4 音声認識タグの定義
Table 4 Definition of speech recognition tags.

音声認識タグ	定義	割合
○	音声認識結果と書き起こし文が完全一致（フィラーが残っている場合は除く）	46%
△	音声認識結果と書き起こし文が意味的に正しい．またはフィラーが残っているがそれ以外は一致	13%
×	誤認識	41%

有名詞が頻出していた．

3.3 音声認識誤りの分類

対話ログのうち日英翻訳が用いられた 84 セッションに含まれる運転手発話について，音声認識誤りの特徴を人手により分析し，6 種類に分類した．6 種類の音声認識誤りの例と相対度数分布を**表 3** に示す．a) 固有名詞の誤認識には地名の誤認識が多く含まれる．b) 同音異義語の誤認識は，たとえば “県庁” から “堅調”，“イカ” から “以下” のような誤りである．c) 似た音素列の単語への誤認識は，たとえば “カニ” から “ハニー” などの誤認識である．d) 発話の一部分の脱落は，発話の一部分が音声認識結果から脱落する誤りである．e) フィラーによる誤認識は，音声認識エンジンの仕様により除去されるべきフィラーが除去されずに認識結果に影響を与えた場合に付与する．f) はその他の音声認識誤りである．分類は発話単位で行い，1 発話に複数の誤りが含まれる場合にはそれぞれの誤りをカウントした．また，固有名詞に対して生じた同音異義語への誤認識や似た音素列への誤認識は，a) としてカウントした．集計の結果，a) 固有名詞の誤認識，b) 同音異義語の誤認識，c) 似た音素列の単語への誤認識，に該当する発話の部分的な置換誤りがすべての誤りの半数以上を占めることが分かった．

また，音声認識誤り識別器の学習用教師ラベルと評価用正解ラベルとするため，発声単位に**表 4** に示すタグを定義した．音声認識誤り識別器の評価実験では ‘○’ と ‘△’ を正例，‘×’ を負例とする．

4. 音声認識誤り識別器

4.1 音声認識誤り識別器の構成

音声認識誤り識別器は認識結果を再検証するポストフィルタである．特定のドメインにそぐわない単語を認識結果が含まないかという観点で，正しい認識結果であるか否かを推定する音声認識誤り識別器を 3 種類検討した．1 つ目は，少量のログをもとに特定ドメインの単語が含まれるか否かを評価できるように，認識結果に含まれる内容語を Bag-of-Words で表現し，素性とする音声認識誤り識別器である．予備実験により，認識結果の形態素数など表層的な 8 つの素性の有効性が確認されたため，Bag-of-Words 以外の素性として追加し，比較用のベースラインとした．2 つ目は音声認識結果の内容語を単語クラス（Bag-of-Word-Classes，BoWC）で表現し，素性とする単語クラスに基づく音声認識誤り識別器である．いずれも識別器はランダムフォレストを使用した．3 つ目は素性として単語分散表現を用いる RNN に基づく音声認識誤り識別器である．学習データに含まれる正例と負例の割合を調整し，負例の検出精度を改善した．なお，Errattahi の手法は，実際には音声認識エンジンが出力する各単語の信頼度をパラメータとして用い，大規模な N-gram コーパスをもとに汎用の単語系列ラベリングを目指したものであるため，特定ドメイン向けに発声単位で識別を行う本研究のベースラインとはしていない．

4.2 Bag-of-Words に基づく音声認識誤り識別器

Bag-of-Words に基づく音声認識誤り識別器は認識結果に含まれる内容語を Bag-of-Words（BoW）で表現し，素性とする．ここで内容語とは，認識結果に対して日本語形態素解析エンジン mecab で形態素解析を行った結果，動詞，名詞，形容詞，副詞の 4 種類とされたものである．素性の一覧を**表 5** に示す．

Bag-of-Words 以外に以下の素性を追加した．まず，認識結果が長いほど音声認識誤りを含む確率が高くなるため，認識結果の形態素数，文字数，音節数など 6 種類の表層的素性を追加した．また，発話の一部分が認識されたときに生じる非文を検出するために，認識結果の先頭と末尾の形

表 5 Bag-of-Words に基づく音声認識誤り識別器の素性と次元数
Table 5 Features and dimensions of speech recognition error identifier based on Bag-of-Words.

素性	次元数	説明
Bag-of-Words	4,882	音声認識結果の内容語の Bag-of-Words 表現
表層的素性	6	音声認識結果の形態素数，文字数，音節数
先頭と末尾の品詞	2	音声認識結果の先頭の形態素の品詞，末尾の形態素の品詞

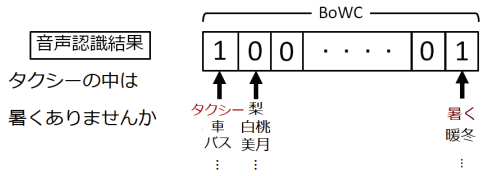

図 1 音声認識結果の Bag-of-Word-Classes 表現
Fig. 1 Representation of speech recognition result in Bag-of-Word-Classes.

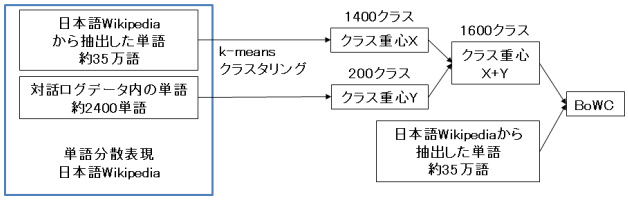

図 2 単語分散表現のクラスタリング手順
Fig. 2 Procedure of clustering word embeddings.

態素が助動詞，格助詞，係助詞，接続助詞であるかを示す品詞情報を 2 種類加えた．識別器にはこれらの素性の組み合わせで判定できるようにランダムフォレストを用いた．

なお，実験において評価データに含まれる内容語も語彙登録しているが，ランダムフォレストの学習データには現れないので，実質的には無視される．

4.3 Bag-of-Word-Classes に基づく音声認識誤り識別器

Bag-of-Word-Classes（BoWC）を用いる単語クラスは，特定環境のドメインで現れる単語と現れない単語を学習データのスパースネスの問題を緩和しながら識別することを意図している．**図 1** に音声認識結果を BoWC で表現する方法を模式的に示す．単語クラスを構成する方法として，シソーラスに基づく方法と単語分散表現のクラスタリングに基づく方法の 2 種類を試した．

4.3.1 シソーラスに基づく Bag-of-Word-Classes

シソーラスは日本語 WordNet [12] を用いた．WordNet は意味辞書であり，名詞，動詞，形容詞，副詞の約 9.3 万単語が概念を表す約 5.7 万種類の synset に分類される．1 つの synset は類義語を纏めたグループであり，is-a 関係で木構造を構成する．木構造の root に近い synset ではより抽象的な意味で，leaf に近い synset ではより具体的な意味で単語が分類される．

単語クラスは以下の方法で構成する．日本語 WordNet の root から一定の深さの階層にあるノードを取り出し，その配下にあるすべての synset に属する全単語を 1 つの単語クラスとする．内容語が複数の synset に属する場合は，そのうちの 1 つの synset をランダムに選択した．WordNet には固有名詞が含まれないため，別途タクシードメインと関連性が高い固有名詞を収集し，固有名詞クラスを追加した．具体的には，インターネット上の観光にかかわるウェブサイトから国内の 8 つの主要な観光地にかかわる文章をそれぞれ 200 万文以上収集し，各観光地を 1 つの文書クラスとして単語の TF-IDF を算出し，TF-IDF を基準にしてタクシー営業地域と関連性が高い単語を抽出した．その単語群から著者ら 3 名の手作業により 1,831 種類の固有名詞を取り出し，さらに対話ログデータの書き起こし文に含まれる固有名詞 191 種類と合わせて，1 つの固有名詞クラスとした．

予備実験の結果，単語クラスを形成するための WordNet の木の階層の root からの深さは 5 とした．その結果，シソーラスに基づく単語クラスは約 7 万語をカバーする 1,686 クラスとなった．なお，synset にない内容語は未知語となり，BoWC 表現に含まれない．

4.3.2 単語分散表現に基づく Bag-of-Word-Classes

単語分散表現に基づく単語クラスは単語分散表現を k-means 法によりクラスタリングしたものである．単語分散表現は単語を数百次元の実数ベクトルで表現したものであり，「単語の意味は，その単語と共起する単語群によって特徴付けられる」という分布仮説に基づいている．単語の意味的な類似性を，ベクトル間の cos 類似度で測ることができる．

単語分散表現は word2vec の skip-gram を用いて獲得した．skip-gram は大規模コーパスと 2 層からなるニューラルネットワークを用いて単語分散表現を獲得する手法であり，特定の単語から周辺単語を予測するタスクを解く形で分散表現を学習する．約 13 億語の日本語 Wikipedia データを用いて，約 45 万語の単語分散表現を得た．そこからタクシードメインに無関係な単語を削減するために，対話ログに含まれる約 2,400 単語との cos 類似度を測り，そのうちの最大値が閾値以上の約 35 万単語に絞った．

さらにドメイン内とドメイン外の単語をできるだけ異なる単語クラスに纏めるため，**図 2** に示す手順で k-means クラスタリングを行った．はじめに日本語 Wikipedia から

表 6 Bag-of-Word-Classes に基づく音声認識誤り識別器の素性と次元数
Table 6 Features and dimensions of speech recognition error identifier based on Bag-of-Word-Classes.

素性	次元数	説明
Bag-of-Word-Classes	1,686/1,600 [1]	音声認識結果の内容語の単語クラス表現
表層的素性	6	音声認識結果の形態素数，文字数，音節数
先頭と末尾の品詞	2	音声認識結果の先頭の形態素の品詞，末尾の形態素の品詞

[1] シソーラスに基づく単語クラスは 1,686 次元，単語分散表現に基づく単語クラスは 1,600 次元とした．

抽出した単語の単語分散表現に対して k-means クラスタリングを行いクラス重心の集合 X を得る．次に対話ログに含まれる単語の単語分散表現に対して k-means クラスタリングを行い，クラス重心の集合 Y を獲得する．両者ともに，日本語 Wikipedia から獲得した共通の単語分散表現を使用する．最後に X と Y を結合したクラス重心の集合を用いてすべての単語分散表現のクラスタリングを行い，単語クラスを作成する．この方法により，ドメイン内の単語が特定のクラスに集中しやすくなると考えた．

予備実験の結果，cos 類似度の閾値を 0.3 とし，単語分散表現に基づく単語クラスは 1,600 クラス，約 35 万単語とした．それ以外の単語は未知語として BoWC 表現には反映されない．BoWC に基づく音声認識誤り識別器の素性の一覧を**表 6** に示す．BoW に基づく音声認識誤り識別器と同様に，認識結果の表層的素性 6 次元と認識結果先頭と末尾の品詞情報の素性 2 次元を加えた．

4.4 Recurrent Neural Network に基づく音声認識誤り識別器

4.4.1 注意機構付き双方向 LSTM-RNN

系列データの長い文脈も考慮できる RNN を用いて音声認識誤り識別器を学習した．対象とする音声認識誤りは，表 3 の a) 固有名詞の誤認識，b) 同音異義語の誤り，c) 似た音素列の単語への誤り，に該当する部分的な置換誤りが半数以上を占めるため，系列中の特定単語に注目できるモデルとして，注意機構付き双方向 LSTM-RNN モデル [13] を適用した．モデルのネットワーク構成を**図 3** に示す．

入力は認識結果を構成する形態素のインデックス系列 $\mathbf{x} = [x_1, x_2, \cdots, x_T]$ である．インデックスは各形態素に対して一意に割り振られた整数である．出力は認識結果に誤認識単語が含まれる確率 y であり，$y \geq 0.5$ の場合に誤認識と判定する．学習時の教師信号は 0/1 の 2 値で与えられる．まず，Input Layer において入力 x_i は one-hot 表現 $\boldsymbol{x}'_i$ に変換され，Embedding Layer において，単語埋め込み行列 $\boldsymbol{E}$ と $\boldsymbol{x}'_i$ との内積により単語分散表現 $\boldsymbol{e}_i$ に変換される．

$$\boldsymbol{e}_i = \boldsymbol{E}\boldsymbol{x}'_i \tag{1}$$

LSTM Layer では $\boldsymbol{e}_i$ を入力としてコンテキストベクトル $\boldsymbol{h}_i$ を得る．LSTM Layer は双方向 LSTM を用いており，順方向 LSTM（forwardLSTM）の出力 $\boldsymbol{fh}_i$ と逆方向

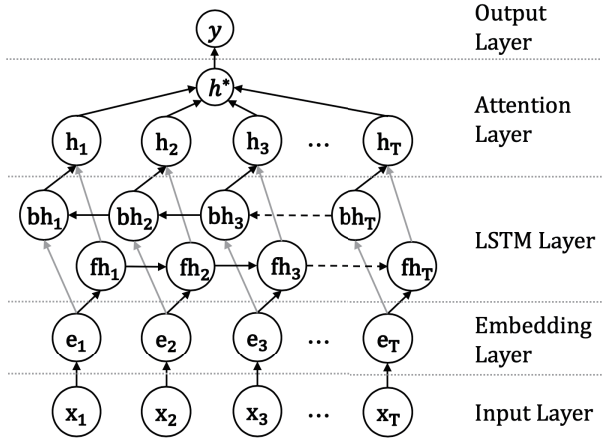

図 3 注意機構付き双方向 LSTM-RNN [13]
Fig. 3 Attention-based Bidirectional LSTM-RNN [13].

LSTM（backwardLSTM）の出力 $\boldsymbol{bh}_i$ の要素和として $\boldsymbol{h}_i$ を得る．

$$\boldsymbol{fh}_i = forwardLSTM(\boldsymbol{e}_i, \boldsymbol{fh}_{i-1}) \tag{2}$$

$$\boldsymbol{bh}_i = backwardLSTM(\boldsymbol{e}_i, \boldsymbol{bh}_{i+1}) \tag{3}$$

$$\boldsymbol{h}_i = \boldsymbol{fh}_i \oplus \boldsymbol{bh}_i \tag{4}$$

Attention Layer では，$\boldsymbol{h}_i$ を列に持つ行列 $\boldsymbol{H}$ と Attention Layer が持つ重みベクトル $\mathbf{w}$ との内積により $\boldsymbol{h}_i\,(1 \leq i \leq T)$ に対する重み係数 a_i を要素とする注視重みベクトル $\boldsymbol{\alpha}$ を求め，$\boldsymbol{h}_i$ に対する加重和により代表ベクトル $\boldsymbol{h}^*$ を獲得する．$\boldsymbol{h}^*$ は全結合の Output layer に入力され，sigmoid 関数を通して出力 y を得る．

$$\boldsymbol{\alpha} = softmax(\mathbf{w}^T tanh(\boldsymbol{H})) \tag{5}$$

$$\boldsymbol{h}^* = tanh(\boldsymbol{H}\boldsymbol{\alpha}^T) \tag{6}$$

$$y = sigmoid(\mathbf{w}_{out}\boldsymbol{h}^* + b_{out}) \tag{7}$$

本モデルは，どのコンテキストベクトル $\boldsymbol{h}_i$ が最終的な出力において重要であるかを教師なしで学習するため，その特徴を誤認識の識別にどの単語が効いているかの推定に利用できると考えた．

予備実験の結果，単語分散表現 $\boldsymbol{e}_i$ を 200 次元，LSTM の各層を 300 次元，コンテキストベクトル $\boldsymbol{fh}_i$，$\boldsymbol{bh}_i$，$\boldsymbol{h}_i$，$\boldsymbol{fh}^*$ を 300 次元とした．Embedding Layer の単語埋め込み行列の初期値は，日本語 Wikipedia をもとに Skip-gram で学習した word2vec で与えた．

4.4.2 学習データにおける正例と負例の調節

表 1 より，今回使用したデータは負例（誤認識の含まれる発話）が正例（誤認識の含まれない発話）に比べて少ない不均衡データであり，そのままの割合で学習データを構成すると正例の検出に偏ってモデルが学習されると考えられる．そこで，学習データに含まれる正例と負例の割合を調整する．具体的には，学習データに含まれる負例をランダムに複製して追加する追加サンプリングを採用する．

5. 評価実験

5.1 評価条件

音声認識誤り識別器の評価実験は，認識結果に誤りが含まれているか否かの 2 値分類である．運転手による 10,019 発話分の認識結果を使用し，10 分割交差検証で評価した．

比較した音声認識誤り識別器は以下に示す A〜F の 6 種類である．識別器 A は素性に Bag-of-Words のみを用い，評価のベースラインとなる．識別器 B〜D はそれぞれ Bag-of-Words，シソーラスに基づく BoWC，単語分散表現に基づく BoWC を基本の素性とし，表層的素性，認識結果の先頭と末尾の品詞情報からなる 8 次元の素性を加えた．識別器 E と F は注意機構付き双方向 LSTM である．識別器 E の正例と負例の比率は 59 : 41 であり，識別器 F では負例の追加サンプリングにより 50 : 50 とした．

A) Bag-of-Words を素性とするRandom Forest
　素性：Bag-of-Words
　識別器：Random Forest
B) Bag-of-Words に素性追加を行ったRandom Forest
　素性：Bag-of-Words + 追加素性 (8)
　識別器：Random Forest
C) シソーラスに基づくBoWC のRandom Forest
　素性：シソーラスに基づく BoWC + 追加素性 (8)
　識別器：Random Forest
D) 単語分散表現に基づくBoWC のRandom Forest
　素性：単語分散表現に基づく BoWC + 追加素性 (8)
　識別器：Random Forest
E) 注意機構付き双方向LSTM
　素性：単語分散表現
　識別器：注意機構付き双方向 LSTM
F) 正例と負例を調節した注意機構付き双方向LSTM
　素性：単語分散表現
　識別器：注意機構付き双方向 LSTM

5.2 負例検出精度・識別精度の評価結果

評価指標は，正例・負例を総合した正解精度（accuracy）に加えて，負例検出の適合率（precision），再現率（recall），F 値（f1 score）とした．各指標の算出式を以下に示す．

$$acc. = \frac{\text{正しく識別した発声の総数}}{\text{発声総数}}$$

$$prec. = \frac{\text{負例と正しく識別された負例の総数}}{\text{負例の総数}}$$

$$recall = \frac{\text{負例と正しく識別された負例の総数}}{\text{識別器が負例として識別した総数}}$$

$$f1 = \frac{2 \times prec \times recall}{prec + recall}$$

表 7 音声認識誤り識別器の識別精度・負例検出精度
Table 7 Evaluation results of the speech recognition error identifiers.

識別器	acc.（%）	prec.（%）	recall（%）	f1（%）
A	65.6	68.4	59.5	57.2
B	69.5	70.2	65.4	65.3
C	69.8	69.9	66.1	66.3
D	70.8	71.3	67.1	67.3
E	74.2	74.7	56.3	64.1
F	75.8	72.9	65.1	68.8

各識別器の正例・負例の識別精度と負例検出精度を**表 7** に示す．識別器 F の正例と負例を調節した注意機構付き双方向 LSTM の正解精度が 75.8%と最も高い．単語クラスに基づく識別器のなかでは D の単語分散表現に基づく BoWC を素性とする識別器の正解精度が最も高く 70.8%であった．

正解精度について符号検定により有意差を検証した．識別器 A と B の間に 1%水準の有意差があった．追加素性は音声認識誤りの特徴をとらえる素性として有効であった．識別器 B と C の間に有意差はなかったが，C と D の間に 1%水準の有意差があった．認識結果を BoW で表現するよりも，BoWC で表現することで学習データのスパースネスの問題が緩和されたと考えられるが，シソーラスに基づく BoWC は単語分散表現に基づく BoWC よりも正解精度は低かった．原因として未知語率の高さが考えられる．単語分散表現に基づく BoWC では未知語率が約 0.1%であるのに対し，シソーラスに基づく BoWC では約 34%と高かった．識別器 D と E の間に 1%水準の有意差があり，注意機構付き双方向 LSTM が有効であった．学習データ量が限られていても，機能語も含めて文脈情報を識別に利用する同手法の方が優れていた．さらに，識別器 E と F の間に 1%水準の有意差があった．負例の追加により負例の再現率が改善されており，正例への学習の偏りが緩和されたと考えられる．

5.3 注意機構の精度評価

注意機構付き双方向 LSTM における注意機構の精度を評価した．注意機構の注視は図 3 の Attention Layer で算出される注視重みベクトル $\boldsymbol{\alpha}$ で表現される．評価データは表 3 の a) 固有名詞の誤認識，b) 同音異義語への誤認識，c) 似た音素列の単語への誤認識，に分類された 740 発話である．

表 8 音声認識誤りラベルと注視ラベルの例
Table 8 Example of speech recognition error labels and attention labels.

書き起こし文	これから	イオン		北	に	向かい	ます.
音声認識結果	これから	家	を	北	に	向かい	ます.
認識誤りラベル	c	c	e	e	c	c	c
注視ラベル	n	n	a	n	n	n	n

表 9 学習データにおける正例・負例の調節前後の注意機構の精度評価
Table 9 Evaluation results of attention mechanism before (E) and after (F) augmentation of negative samples.

識別器	閾値	prec.（%）	recall（%）	f1（%）	acc.（%）
E	0.1	54.5	42.8	47.9	80.0
	0.2	60.5	38.2	46.8	81.4
F	0.1	59.5	46.0	51.9	81.7
	0.2	64.5	41.1	49.8	82.2

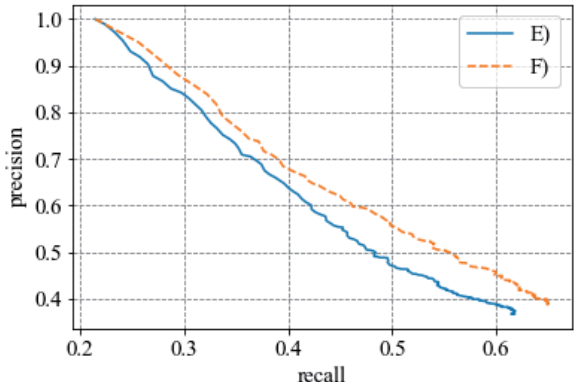

図 4 学習データにおける正例・負例の調節前後の注意機構の Precision-Recall 曲線
Fig. 4 Precision-Recall curves of attentions before (E) and after (F) augmentation of negative samples.

精度評価は，認識結果の形態素列に対する以下の 2 種類のラベルの比較に基づく．1 つは形態素が正しく認識された (c) か誤って認識された (e) かを示す認識誤りラベル，もう 1 つは識別器により注視された (a) か注視されていない (n) かを示す注視ラベルである．注視ラベルは，基本的に Attention Layer の重み係数に対する閾値処理に基づくが，閾値以上の重み係数が存在しない音声認識結果については，重み係数最大の形態素を注視された (a) とした．ラベルの例を**表 8** に示す．

注視の評価指標は以下の 4 種類である．

$$acc. = \frac{(\text{認識誤り e かつ注視 a}) \text{ または } (\text{c かつ n}) \text{ の総数}}{\text{音声認識結果に含まれる形態素の総数}}$$

$$prec. = \frac{\text{認識誤りラベル e かつ注視ラベル a の形態素総数}}{\text{注視ラベル a の形態素の総数}}$$

$$recall = \frac{\text{認識誤りラベル e かつ注視ラベル a の形態素総数}}{\text{認識誤りラベル e の形態素の総数}}$$

$$f1 = \frac{2 \times prec. \times recall}{prec. + recall}$$

閾値を 0.1，0.2 とした場合の，識別器 E と F における注視の精度を**表 9** に示す．注意機構の教師信号を与えていないが，80%を超える正解精度が得られている．また，閾値を 0 から 1 まで変化させたときの Precision-Recall 曲線を**図 4** に示す．閾値の設定にかかわらず，学習データにおける正例と負例の割合を調節した方が調整前よりも負例検出の精度が優れた．

6. 考察

音声認識誤り識別器は，基本的に正しい認識結果の誤棄却を小さな割合に抑えたうえで，誤認識の検出率を高めるのが適切と考えられる．評価実験では識別精度（accuracy）を最優先の指標として，次に誤認識の検出精度を評価した．

ランダムフォレストに基づく識別器において，表層的素性，認識結果の先頭と末尾の形態素の品詞情報の計 8 次元の素性追加は顕著に有効であった．BoW に較べて，単語クラス（BoWC）の導入は有効であったが，シソーラスに基づく BoWC による改善度は小さかった．未知語による情報の欠落の影響が大きいと考えられる．

注意機構付き双方向 LSTM に基づく識別器は文のコンテキストを考慮することができるが，ドメインの学習データは約 1 万文と限られており，過学習が発生していると予想される．それでも，注意機構付き双方向 LSTM に基づく識別器が優れた識別精度を示した．評価ドメインのデータは極端な正例・負例の不均衡を示してはいないが，負例の追加サンプリングにより負例検出の F 値は最高の 68.8%を示し，全体の識別精度改善にもつながった．

7. おわりに

音声認識技術や機械翻訳技術の高精度化に合わせて，今後は汎用的な音声認識エンジンをカスタマイズして様々な機能と組み合わせ，特定用途向けの音声サービスが立ち上げられると予想される．このような状況を想定し，本研究では音声認識エンジンをブラックボックスとしてとらえ，音声認識結果が特定ドメインの語彙以外の誤認識単語を含むかどうかを，特定ドメインの限られた量のログデータをもとに識別する音声認識誤り識別器を検討した．

音声認識結果に含まれる内容語の BoW 表現を特徴量とするランダムフォレスト識別器をベースラインとして，特徴量のデータスパースネスを緩和するために，特徴量を単語クラス（BoWC）に代替する方法を提案した．BoWC の構成方法として，シソーラスに基づく方法と単語分散表現のクラスタリングに基づく方法を比較評価し，単語分散表現のクラスタリングに基づく方法の方が高い精度が得られることを確認した．加えて，追加素性として，発話を構成する形態素の数，文字数，音節数などの表層的素性と認識結果の先頭と末尾の形態素の品詞情報が有効であることを確認した．

また，文のコンテキストを考慮できる RNN に基づく識

別器として，誤り単語に動的に焦点を当てられる特徴をもつ注意機構付き双方向 LSTM モデルを適用した．正解精度を高く保ちながら，負例の検出率を改善するために学習データにおける正例と負例の割合を調整した．

特定ドメインの例としてタクシー車内音声翻訳システムの試験サービスで収集された約 1 万発声分の音声認識結果テキストを用いて 3 種類の音声認識誤り識別器を学習し，比較評価した．その結果，学習データにおける正例と負例の割合を調整した注意機構付き双方向 LSTM が最も高い識別精度（75.8%）を示し，教師なしで学習された注意機構の注視精度は 80%を超えることを確認した．

参考文献

[1] 松田繁樹，林 輝昭，韋苅 豊，志賀芳則，柏岡秀紀，安田圭志，大熊英男，内山将夫，隅田英一郎，河井 恒，中村 哲：多言語音声翻訳システム “Voice Tra” の構築と実運用による大規模実証実験，電子情報通信学会論文誌，Vol.J96-D, No.10, pp.2549–2561 (2013).

[2] Kemp, T. and Schaaf, T.: Estimating confidence using word lattices, *Proc. Eurospeech 1997*, pp.827–830 (1997).

[3] Wessel, F., Schluter, R., Machery, K. and Ney, H.: Confidence measures for large vocabulary continuous speech recognition, *IEEE Trans. Speech and Audio Processing*, Vol.9, No.3, pp.288–298 (2001).

[4] Sukkar, R.A. and Lee, C.-H.: Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition, *IEEE Trans. Speech and Audio Processing*, Vol.4, No.6, pp.420–429 (1996).

[5] Rahim, M.G., Lee, C.-H. and Juang, B.-H.: Discriminative utterance verification for connected digits recognition, *IEEE Trans. Speech and Audio Processing*, Vol.5, No.3, pp.266–277 (1997).

[6] Tam, Y.C., Lei, Y., Zheng J. and Wang, W.: ASR error detection using recurrent neural network language model and complementary ASR, *Proc. ICASSP 2014*, pp.2312–2316 (2014).

[7] Ogawa, A., Hori, T. and Nakamura, A.: Error type classification and word accuracy estimation using alignment features from word confusion network, *Proc. ICASSP 2012*, pp.4925–4928 (2012).

[8] Korenevsky, M.L., Smirnov, A.B. and Mendelev, V.S.: Prediction of speech recognition accuracy for utterance classification, *Proc. Interspeech 2015*, pp.1275–1279 (2015).

[9] Ogawa, A. and Hori, T.: Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks, *Speech Communication*, Vol.89, pp.70–83 (2017).

[10] Errattahi, R., Hannani, A.E., Hain, T. and Ouahmane, H.: Towards a generic approach for automatic speech recognition error detection and classification, *Proc. Advanced Technologies for Signal and Image Processing* (2018).

[11] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J.: Distributed representations of Words and phrases and their compositionality, *Proc. Neural Information Processing System 26*, pp.3111–3119 (2013).

[12] Bond, F., Isahara, H., Uchimoto, K., Kuribayashi, T. and Kanzaki, K.: Extending the Japanese WordNet, *15th Annual Meeting of The Association for Natural Language Processing*, C1-4, pp.80–83 (2008).

[13] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H. and Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification, *Proc. Annual Meeting of the Association for Computational Linguistics 2016*, pp.207–212 (2016).

田中 拓

2019 年同志社大学理工学部インテリジェント情報工学科卒業．同年オムロンソフトウェア株式会社入社．ニューラルネットワーク技術に関心．

野田 直希

2017 年同志社大学理工学部インテリジェント情報工学科卒業．2019 年同大学大学院情報工学専攻博士前期課程修了．同年富士通株式会社入社．音声認識等の誤り検出・訂正技術に関心．

加藤 恒夫 （正会員）

1996 年東京大学工学系大学院・電子工学専攻博士前期課程修了．同年，国際電信電話株式会社入社．KDD 研究所，KDDI 研究所を経て，2015 年より同志社大学理工学部インテリジェント情報工学科准教授．音声認識，ユーザインタフェースの研究開発に従事．2011 年情報処理学会喜安記念業績賞受賞．電子情報通信学会，日本音響学会，言語処理学会，ヒューマンインタフェース学会，ACM，IEEE 会員．博士（情報理工学）．本会シニア会員．

山本 誠一 （正会員）

昭和 25 年生．昭和 47 年大阪大学工学部電子工学科卒業．昭和 49 年同大学院基礎工学研究科修士課程終了．同年国際電信電話株式会社入社．ATR 音声言語コミュニケーション所長等を経て現在，同志社大学理工学部教授．工博．この間，適応信号処理，音声合成，音声認識，音声翻訳，マルチモーダル対話の研究等に従事．1981 年度電子情報通信学会学術奨励賞，日本音響学会第 3 回技術開発賞，第 5 回技術開発賞，電子情報通信学会システムソサイエティ論文賞，電気通信普及財団テレコム技術賞等を受賞．日本音響学会，言語処理学会，人工知能学会各会員，IEEE Fellow，電子情報通信学会 Fellow．