E-018

国籍情報を用いた人名の音訳
Improving Transliteration Quality of People's Name by Considering Nationality Information

宮崎 太郎† 熊野 正† 今井 篤†
Taro Miyazaki Tadashi Kumano Atsushi Imai

1. はじめに

放送局では視覚障害者に向けた情報提供として解説放送のサービスを実施している．放送音声だけでは伝わりづらい人物の表情などを音声で提供するもので，ドラマなどの一部のコンテンツに付与されている．しかし，制作コストが高いことから，放送時間のうち 10%程度の番組にしか付与できていない[1]．特にスポーツ中継などの生放送番組では，音声で伝えられていない情報を見つけ，それをどう伝えるかを決定し，それを伝えるという作業を瞬時に行う必要があり，作成が困難である．一方で，スポーツ中継は人気の高いコンテンツであり，映像と音声だけでなく，字幕や解説放送を付与することが望まれている．これが実現できれば，より多くの人が一緒に楽しめるコンテンツになる．

NHK では，視覚障害者に向けた新たなサービスとして，特にスポーツ中継を対象とした音声ガイドの自動生成についての検討を開始している[2,3]．音声ガイドでは，スポーツのメタデータを解析し，その時点で何が起きているのかなどの情報を自動音声で伝える．これにより，視覚障害者への情報バリアフリーを実現することを目指している．

音声ガイドを実現するうえでの課題の一つに，人名の読みカナ付与がある．ロンドン五輪に出場した約 60,000 人の選手のうち，NHK が事前にカナ表記が定めていたのは 14,000 人程度であった．音声ガイドの実現には，それ以外の選手について，アルファベット表記からカナ表記を自動生成することが必要である．

本稿では，人名のアルファベット表記からカナ表記を自動生成する手法について述べる．人名は，例えば同じ「Michael」という表記でも，英語圏であれば「マイケル」，フランス語圏であれば「ミシェル」，ドイツ語圏であれば「ミヒャエル」などと，言語に応じて読みが異なる場合がある．提案手法では，その人物の国籍情報を入力に加えることで，言語による表記異なりも表現できることを目指す．

2. 提案手法

提案手法ではニューラルネット(NN)を用いる．NN を用いた機械翻訳は，Encoder-Decoder モデル[4]と呼ばれる，まず翻訳元の文章を再帰型ニューラルネット(RNN) に入力(Encode)し，すべて入力したらそれを翻訳先の言語に翻訳(Decode)する手法が主流である．翻訳元の文章の意味を RNN が解釈し，その意味に基づいた翻訳先の文章を生成することができるため，語順の入れ替えを含めた機械翻訳に有効である．

一方で，音訳は語順の入れ替えが発生しない．そのため，一度 Encode する必要がなく，原言語の文字を前から順に解釈し，その都度対応するカナを出力することで実現できる． 提案手法では，RNN を用い，アルファベット表記の入力を前から順にカナに置き換える．

†NHK 放送技術研究所
NHK Science & Technology Research Laboratories

2.1 RNN を用いた人名の音訳

RNN を用いた人名の音訳は，アルファベット表記の入力を前から順に RNN に入力し，その時点で出力できるカナがあれば出力し，出力するものがなければ φ を出力する．図 1 に，入力が「Steve」の場合の例を示す．入力側の「<s>」「<e>」はそれぞれ入力の先頭，終了を示す記号である．このように，RNN の入力に音訳したい文字列を 1 文字ずつ入力し，RNN がそれを受け取るごとにそれに対応するカナ，または φ を出力する．

RNN のモデル学習には， Connectionist Temporal Classification (CTC)[5]を用いて誤差計算し，その結果をによる誤差逆伝搬法を用いる．CTC は入力と正解データの系列長が異なる場合に，正解データの系列に空白文字を挿入し，正解データが正しい順序で出力される確率を求める．音訳のように，入力の複数の文字をまとめて一つの出力を得る必要があるときに有効である．

なお，本稿では，促音（「ッ」）はそれに続くカナとまとめ，長音記号（「ー」）と拗音（「ャ」，「ュ」など）はその前のカナとまとめて一つのトークンとして扱う．表 1 にトークンの例を示す．

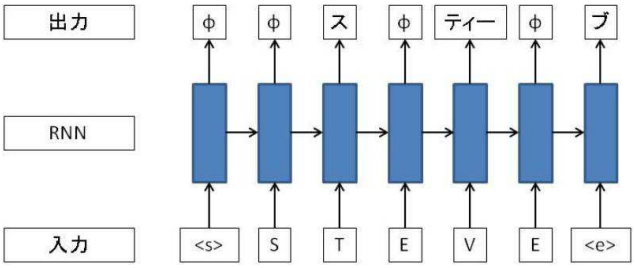

図 1 RNN を用いた音訳

表 1 トークンの例

ア	カナ	ット	促音＋かな
ジョ	カナ＋拗音	ッジョ	促音＋カナ＋拗音
ター	カナ＋長音	ッター	促音＋カナ＋長音
ツァー	カナ＋拗音＋長音	ッツァー	促音＋カナ＋拗音＋長音

2.2 国籍情報を用いた人名の音訳

2.1 節で述べた音訳手法では，入力されたローマ字の文字列のみを手がかりとして，カナ表記を推定する．この手法では同一の表記でも言語により読み方が異なる場合に，言語に応じた訳し分けができない．そこで，RNN にその人物の国籍を入力することで，人物の国籍に応じた訳し分けを目指す．国籍は必ずしもその選手の名前の言語を表すわけではないが，その手がかりになることから，音訳の性能の向上が期待できる．

国籍情報を導入した固有名詞の音訳手法の概要を図２に示す．2.1 節で述べた手法との違いは，先頭記号<s>を RNN に入力した直後に，国籍を入力する点である．これにより，それ以降に入力されるローマ字の文字列を音訳する際に国籍情報を加味したカナ表記を出力することが可能となる．

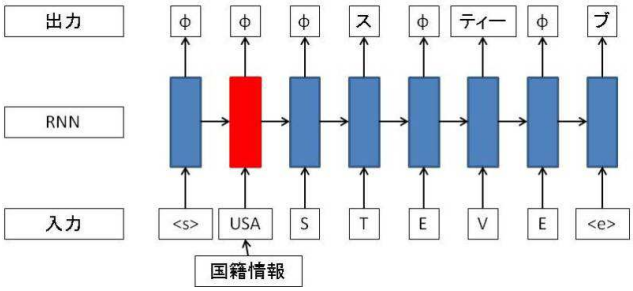

図２　国籍情報を用いた RNN による音訳

3. 評価実験

3.1 実験条件

評価実験では，2.1 節，2.2 節で述べた２つの手法と，２つのベースライン手法を比較した．ベースライン手法には，フレーズベース統計的機械翻訳手法を用いた音訳と，RNN を用いた Encoder-Decoder モデルを用意した．Encoder-Decoder モデルの概要を図３に示す．

RNN を用いた手法の実装には Chainer[1]を，フレーズベース統計的機械翻訳には Moses[2]を用いた．RNN を用いたモデルでは，すべて中間層の数を 1,000 個とした．

ロンドン五輪に参加した選手で人手によるカナ表記が付与されている 13,650 人分のうち，無作為に抽出した 1,000 人分を正解データ，残りの 12,650 人分を学習データとした．

RNN では学習時にランダム要素を含むため，学習ごとに異なったモデルができる．そこで，RNN を用いた手法では同一の条件で３回のモデル学習を行った．各回の epoch ごとの誤差を比較し，学習時の誤差が最小となる epoch のモデルを用いて評価した．

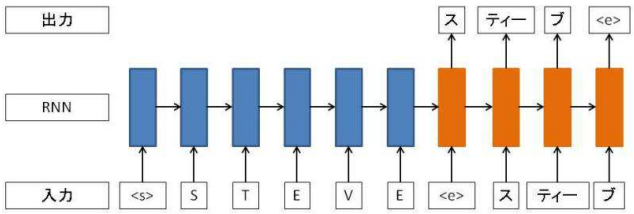

図３　Encoder-Decoder モデルの概要

3.2 評価実験結果

表 2 に評価実験の結果を示す．表中の PBSMT，Encoder-Decoder はそれぞれベースライン手法のフレーズベース統計的機械翻訳による音訳，Encoder-Decoder モデルによる音訳を表し，RNN は 2.1 節で述べた RNN を用いた音訳，RNN+国籍情報は 2.2 節で述べた国籍情報を用いた音訳を表す．正解率は音訳が完全に正解データと一致した場合を正解とした音訳の正解率を表し，BLEU はトークン単位で計算した BLEU 値である．

提案手法である RNN+国籍情報がいずれの指標においても最も高い性能を示した．RNN による手法は，国籍情報を用いない場合でも Encoder-Decoder モデルによるものを上回っており，音訳の場合には入力を前から順番に変換できる RNN による手法を用いたほうが良いことがわかる．

フレーズベース統計的機械翻訳による手法は，国籍情報を用いない場合の２つの手法と比較して高い性能を示したが，この手法に国籍情報を導入することは難しい．

同じ表記でも国籍により異なるカナを付与する場合の例を表３に示す．この例では，Peter という表記で，国籍の異なる２名の名前を翻訳した．Peter は英語では「ピーター」，ハンガリー語では「ペーテル」と発音する．国籍情報を用いて音訳する提案手法ではこの訳し分けができている． それに対し，国籍情報を用いない手法ではそのような区別をできないため，異なる言語での発音が混ざったような翻訳結果が出力された．

表２　評価実験結果

手法	正解率	BLEU
PBSMT	18.6%	55.57
Encoder-Decoder	10.5%	47.14
RNN	13.4%	56.30
RNN＋国籍情報	**20.7%**	**63.54**

表３　同表記で異なるカナを付与する例

アルファベット表記	Peter Taylor（アイルランド）	Peter Kusztor（ハンガリー）
カナ表記	ピーター・テイラー	ペーテル・クストル
RNN	ペタール・タイラー	ペタール・クストル
RNN +国籍情報	**ピーター**・テイラー	**ペーテル**・クストル

4. おわりに

本稿では，人名を対象とした音訳手法について述べた．RNN を用いた音訳に，国籍情報を入力として加えることで，音訳の正解率が 20.7%となり，フレーズベース統計的翻訳手法を用いたベースラインと比較して 2.1 ポイント，国籍情報を考慮しない RNN による音訳手法と比較して 7.3 ポイントの精度向上が得られた．

今後の課題として，アメリカのように国籍情報だけでは名前の言語の推定に不十分な国について，それに対応するための手法の検討が必要である．

参考文献

[1] 総務省，“平成 26 年度の字幕放送等の実績，” http://www.soumu.go.jp/menu_news/s-news/01ryutsu09_02000126.html
[2] 今井 篤 他,”テレビ番組へのオーバーラップを許容した音声補助情報サービスの検討,” 電子情報通信学会総合大会, H-4-11, pp. 322 (2016).
[3] 佐藤 庄衛 他, “スポーツ中継における音声ガイドの有効性の調査,” 日本音響学会 2016 年春季研究発表会, 1-4-7, pp.1547-1548 (2016).
[4] Ilya Sutskever et al., “Sequence to Sequence Learning with Neural Networks,” NIPS (2014).
[5] Alex Graves et al., “Connectionist temporal classification: labeling unsegmented sequence data with recurrent neural networks," *the 23rd international conference on Machine learning*. pp. 369-376 (2006).

[1] http://chainer.org/
[2] http://www.statmt.org/moses/