音声認識単語仮説の曖昧性を考慮するニューラル機械翻訳

長村 佳歩　　叶 高朋　　Sakriani Sakti　　須藤 克仁　　中村 哲
奈良先端科学技術大学院大学 情報学研究科
{osamura.kaho.oe5, kano.takatomo.km0, ssakti, sudoh, s-nakamura}@is.naist.jp

1 はじめに

音声翻訳システムは，音声認識 (ASR) と機械翻訳 (MT) の最低 2 つのシステムで構成されるシステムである．音声翻訳システムでは，入力された音声はまず ASR によってテキスト化され，次に MT によって目的の言語に翻訳される．音声翻訳において音声認識誤りがあれば，誤った入力を受ける MT の出力も間違う可能性が高い．音声認識技術は向上を続けているものの音声認識誤りの発生は不可避であり，それを無視して音声翻訳システムを構築すると実用性を欠いてしまう．そこで従来研究で提案されたのが，原言語側に音声認識誤りを含む対訳文を MT の学習に利用する手法である [1, 2, 3]．これらの手法により，音声認識誤りに頑健な翻訳が可能になったが，複雑なモデルが必要であったり，MT の学習に多くの音声データが必要など改良の余地があった．

本研究では，ニューラルネット機械翻訳 (NMT) を用いた音声認識誤りに頑健な音声翻訳を目指した．原言語側に音声認識誤りを含む入力を NMT の学習に利用するために，通常 NMT では one-hot ベクトルを入力とするところを，音声認識候補の曖昧性を表現するベクトルを入力として用いる．提案手法では，テキストの翻訳時の単語表現を用いて音声認識候補を表現しているため，テキストからの学習と音声認識結果からの学習を同じ枠組みで行うことができる．テキスト音声合成 (TTS) による合成音声を用いたシミュレーション実験において，提案手法は音声認識の 1-best を翻訳するベースラインシステムと比較して BLEU が 4.8 から 5.8 向上し，提案手法が音声認識誤りに頑健な翻訳を可能にすることを示した．

2 関連研究

大串ら [2] は統計的機械翻訳 (SMT) を用いた音声翻訳において，音声認識候補の N-best と ASR で用いられる特徴を使用して，N-best からの翻訳精度を基準にした単語選択を行うことで，音声認識誤りに対応した．ASR は認識確率最大の単語を出力するように学習されており，MT は翻訳確率最大化を目標として翻訳を行っている．そのため，ASR の出力が MT に適した出力ではない場合があった．そこで，音声認識候補から翻訳指標に基づいた単語の選択を行い，音声認識誤りに対応した音声翻訳を実現した．本研究でも，音声認識候補の N-best を用いているが単語選択は行わず，N-best の情報を全て用いている．また，各認識候補の曖昧性を表現するベクトルとして N-best を MT に学習させている．さらに，大串らの研究では SMT が用いられているが，本研究では NMT を用いた．

Sperber ら [3] は NMT を用いた音声翻訳において，音声認識結果の音声言語構造の情報を表現したグラフ構造 (単語ラティス) と各パスの信頼度を用いて翻訳することで，音声認識誤りに頑健な翻訳を行った．この研究では，NMT を単語ラティス入力に対応するように拡張するための LatticeLSTM を考案し，入力の曖昧性を考慮した翻訳を実現している．しかしながら通常の曖昧性のない NMT と比較して計算が複雑になるという問題がある．本研究は単語ラティスではなく，Word Confusion Networks(WCNs)[9] のように対立単語仮説の事後確率分布を利用した．WCNs は単語ラティスに比べ表現可能な曖昧性が各単語仮説に対する対立候補に限定されるものの，各時点での単語仮説の曖昧性を直接的に表現可能である．本研究ではこの単純な構造を利用することで，NMT そのものの計算方法を大きく変更することなく入力の曖昧性を考慮することができる手法を提案する．

3 ニューラルネットを用いた音声認識と機械翻訳

本研究で用いる英日音声翻訳システムについて説明する．音声翻訳システムは ASR と NMT で成り立っており，どちらもニューラルネットワークを用いたものとなっている．ここで，本研究でのベースラインと

なる Bahdanau ら [4] が提案した注意型エンコーダ-デコーダモデルに基づく ASR と NMT について説明する．

3.1 注意型エンコーダ-デコーダモデル

注意型エンコーダ-デコーダモデルとは入力系列から目標の出力系列を得るためのニューラルネットワークを用いたモデルである．モデルでは，入力シンボルの系列 $\boldsymbol{x} = x_0 \cdots x_{|x|}$ に対して，出力シンボルの系列 $\boldsymbol{y} = y_0 \cdots y_{|y|}$ を求める．このシステムでは入力シンボルをエンコードするエンコーダと，デコーダにエンコーダの情報を与えるアテンション，出力シンボルを生成するデコーダからなっている．

エンコーダでは入力シンボル $\boldsymbol{x}$ が与えられたとき，LSTM[5] を用いてベクトルに変換する．j 番目のシンボルでのエンコーダの隠れ層は前向きの LSTM と後ろ向きの LSTM を用いて，

$$h_j = [\overrightarrow{h_j}^{\mathrm{T}} : \overleftarrow{h_j}^{\mathrm{T}}]^{\mathrm{T}} \tag{1}$$

を作成する．$\overrightarrow{h_j}$ 及び $\overleftarrow{h_j}$ は，それぞれ

$$\overrightarrow{h_j} = \mathrm{LSTM}(x_j, h_{t-1}), \overleftarrow{h_j} = \mathrm{LSTM}(x_j, h_{t+1}) \tag{2}$$

と計算される．

デコーダは入力シンボル系列 $\boldsymbol{x}$ が与えられた場合の，出力シンボル系列 $\boldsymbol{y}$ の確率を計算する．出力シンボル系列の確率を個々の出力シンボルの確率の累積に分解して計算して，出力シンボル系列 $\boldsymbol{y}$ を得る．i 番目の出力シンボルの事後確率は LSTM を用いて，

$$p(\hat{y_i}|\boldsymbol{y}_{<i}, \boldsymbol{x}) = \mathrm{softmax}(\mathrm{LSTM}(s_i, y_{i-1}, c_i)) \tag{3}$$

として計算される．デコーダでの隠れ層 s_i は LSTM をもちいて，

$$s_i = \mathrm{softmax}(\mathrm{LSTM}(s_{i-1}, y_{i-1}, c_i)) \tag{4}$$

として，前ステップの隠れ層 s_{i-1} と出力シンボル y_{i-1}，アテンションベクトル c_i を用いて計算される．

アテンションで用いられるアテンションベクトル c_i は，エンコーダの隠れ層 h_j の重み付き和であり，

$$c_i = \sum_{j=1}^{|x|} \alpha_{ij} h_j \tag{5}$$

で表わされる．式 5 で示される α_{ij} は，softmax 関数を用いて総和が 1 になるように正規化されている．α_{ij} は，

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{|x|} \exp e_{ik}} \tag{6}$$

$$e_{ij} = v^{\mathrm{T}} \tanh(W_s s_{i-1} + W_h y_j) \tag{7}$$

ここで，v は重みベクトル，W_s と W_h は重み行列である．α_{ij} がアテンション確率を表わしており，y_i が x_j から受ける影響を確率的に表現しているとみなすことができる．

3.2 音声認識

音声認識は，音声の特徴量が与えられたときに，その単語列を推定する問題である．その為，音声認識では入力が音声の特徴量系列 $\boldsymbol{x_a} = x_{a_0} \cdots x_{a_{|x_a|}}$ となり，出力を単語系列 $\boldsymbol{y_a} = y_{a_0} \cdots y_{a_{|y_a|}}$ とする．ここでは，y_{a_i} は単語の分散表現である．

3.3 機械翻訳

機械翻訳は，原言語の単語系列が与えられたときに，目的言語の単語系列を推定する問題である．その為，機械翻訳では入力が原言語の単語系列 $\boldsymbol{x_m} = x_{m_0} \cdots x_{m_{|x_m|}}$ となり，出力を目的言語の単語系列 $\boldsymbol{y_m} = y_{m_0} \cdots y_{m_{|y_m|}}$ とする．ここでは，x_{m_j} と y_{m_i} は単語の分散表現である．

4 提案手法：音声認識候補の曖昧性を表現するベクトルを用いた翻訳

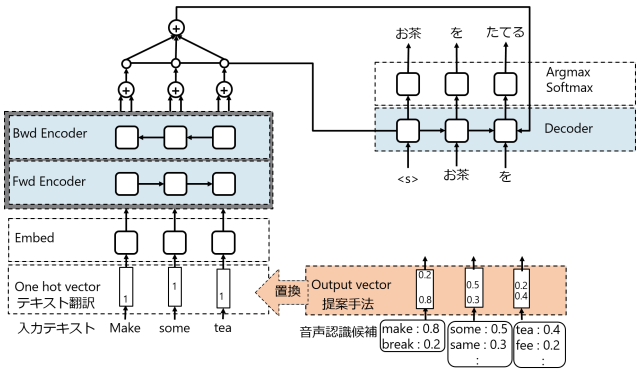

図 1: 提案手法とテキスト翻訳の差分

本研究では，MT に音声認識候補の曖昧性を表現する正規化されたベクトル (output ベクトル) を与えることで，音声認識誤りがある場合であっても最適な文に翻訳することを目指す．output ベクトルは one-hot ベクトルと次元数が同じで，各次元が単語の事後確率を表す．output ベクトルでは，WCNs のように対立単語仮説の事後確率分布を利用しているため，入力された単語だけでなく似た音の単語の事後確率も高くなる傾向がある．例えば”make”の音声が入力されたとき，”make”の事後確率だけでなく似た音の単語である”break”の事後確率も高くなる．

図 1 に，テキスト翻訳の NMT と提案手法の差分を示した．テキスト翻訳の NMT では入力された単語を one-hot ベクトルに変換し，モデルの次元数へと線形変換を行って単語の分散表現を作成した．しかし，本手法では，学習時に one-hot ベクトルを用い，チューニングで output ベクトルを用いた．チューニングでは，output ベクトルに対し，one-hot ベクトルと同じ処理を行い単語の分散表現を作成した．output ベクトルの各要素に対応する語を one-hot ベクトルと等しくしているため，テキストからの学習と音声認識結果からの学習を同じ枠組みで行うことが可能である．

ここで，output ベクトルには音声認識候補と式 3 の事後確率を用いる．確率が高い単語は単語は ASR の入力として与えられた可能性が高く，確率の低い単語は入力として与えられた可能性は低い．音声認識では式 3 の結果から，もっとも確率の高い単語 y_{a_i} を出力していたが，本研究では，音声認識の入力特徴が x_{a_j} の時には，$p(\boldsymbol{y_a}|x_{a_j})$ を MT の入力とする．

5 実験

提案した，音声認識候補の曖昧性を表現する正規化されたベクトルを用いた翻訳の精度を確認するため，英日翻訳のタスクで実験を行った．今回行った実験は音声翻訳のテストや ASR の学習で合成音声を利用したシミュレーション実験である．

5.1 データセットと評価指標

実験で用いたデータセットは旅行会話コーパス BTEC[7] である．ASR の学習では，Google Text to Speech (gTTS) API[6] を用いて，159k 文のテキストデータを音声へ変換したデータを用いた．また，音声のサンプリング周波数は 16kHz，23 チャネルのメルフィルタバンクをもちいた．NMT の学習では英語と日本語のパラレルコーパス 464k 文を用いた．また，チューニングでは，学習時のコーパスから選択した 159k 文から gTTS をもちいて生成した音声を使用している．output ベクトルを用いた翻訳のために，NMT はテキストの各単語を one-hot ベクトルに変換して学習時の入力とし，output ベクトルによってチューニングを行った．さらに，ASR と MT のテストでは 508 文のパラレルコーパス用いている．TTS 生成の音声を ASR の入力とし，ASR のモデルの出力する output ベクトルを NMT の入力としてテストを行った．

ASR と NMT のモデルには LSTM を用い，エンコーダでは双方向の LSTM を用いている．パラメータは語彙数が 16745 語 (ASR,NMT 共通)，エンコーダの次元数 500，隠れ層の次元数 500，ドロップアウトを 0.1，バッチサイズを 32 とし，最適化手法には Adam(学習率エンコーダ：0.0001，デコーダ：0.0005) を用いた．ASR のモデル学習時の epoch 数は 4,11,15,74，NMT のモデル学習時の epoch 数は 120 であり，提案手法の output ベクトルによるチューニングは 2〜4epoch である．今回の実験は学習時の epoch が異なる ASR のモデルを用いる．学習時の epoch 数の違いでモデルを分けており，全てのモデルのパラメータは等しい．epoch 数が 4 のモデルの WER は 15.17%であり，epoch 数 11 は WER=12.34%，epoch 数 15 は WER=11.05%，epoch 数 74 は WER=8.82%である．

5.2 モデル

実験にはベースラインとして，ASR の 1-best を NMT の入力として翻訳するベースラインモデルを用意した．また，提案手法として，output ベクトルを NMT 入力とする提案手法モデルを用い．さらに，比較基準として ASR の参照文を入力したテキスト翻訳結果を載せている．ベースラインとテキスト翻訳のモデルは one-hot ベクトルで学習しており，提案手法のモデルは one-hot ベクトルで学習後，ASR から output ベクトルでチューニングを行った．チューニングでは 5.1 節にある ASR の各モデルを用いている．

6 結果

実験結果を BLEU[8] を用いて評価し，図 2 に示した．提案手法によって，ベースラインモデルと比較して BLEU が 4.8 から 5.8 向上した．

図 2 の Proposed は提案手法，Baseline はベースライン，Reference はテキスト翻訳の結果をそれぞれ示しており，横軸のラベルはチューニング時に用いた ASR のモデルの WER を示している．

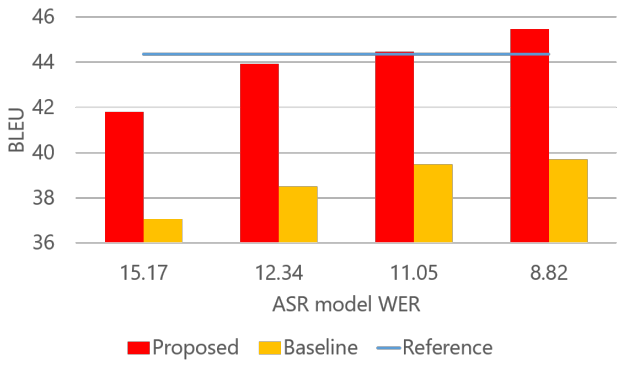

図 2: 各モデルの BLEU

7 考察

英日翻訳における，ベースラインと提案手法の出力例を表1に示す．ASRにおいて”shoe”が”station”と

表 1: ベースラインと提案モデルの出力例 (epoch=74のモデルを使用)

ASR 参照文	excuse me where is the closest shoe store
ASR 認識結果	excuse me where is the closest station store
Baseline	すみません一番近い駅はどこですか
Proposed	すみません一番近い靴屋はどこですか
MT 参照文	すみません一番近い靴屋はどこですか

音声認識誤りされたために，ベースラインでは”靴屋”でなく”駅”と翻訳されてしまっている．しかし，提案手法では”靴屋”と翻訳出来ており，これはoutputベクトルに含まれる”shoe”の情報を用いて翻訳を行えたためだと考えられる．

しかし，常に表1のようになるのではなく，ベースラインよりも提案手法が翻訳誤りをしている場合や，ベースラインと同じ誤りをしている翻訳も存在する．チューニングに用いたASRのモデルのWERが上昇すればするほど，翻訳誤りの傾向は強くなる．だが，テスト全体の精度はベースラインをしのいでおり，提案手法によってより最適な翻訳をより行うことができたと考えられる．また，図2においてテキスト翻訳を超える精度が得られたのは，outputベクトルに含まれる認識候補から，パラフレーズの情報を得ることができたためである可能性が考えられる．

8 おわりに

本研究ではNMTへの入力に対し，音声認識候補の曖昧性を表現する正規化されたベクトルであるoutputベクトルを与えることで，音声認識誤りに頑健な音声翻訳システムの実現をおこなった．結果として，outputベクトルによってベースラインよりBLEUが4.8から5.8向上した．音声翻訳において，音声認識候補を用いることで，NMTにとって最適な音声認識候補の選択ができた可能性がある．

謝辞

本研究の一部はJSPS科研費 JP17H06101，JP17K00237の助成を受けたものです．

参考文献

[1] Nicholas Ruiz, Qin Gao and William Lewis et al. Adapting Machine Translation Models toward Misrecognized Speech with Text-to-Speech Pronunciation Rules and Acoustic Confusability. INTERSPEECH, 2015, pp.2247-2251.

[2] 大串正矢, Graham Neubig and Sakriani Sakti et al. 音声認識と機械翻訳のランク学習による同時最適化言語処理学会, 2013.

[3] Matthias Sperber, Graham Neubig and Jan Niehues et al. Neural Lattice-to-Sequence Models for Uncertain Inputs. EMNLP, 2017.

[4] Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. ICLR, 2015.

[5] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. Neural Computation, Vol.9, No.8, pp.17351780, 1997.

[6] Google Text to Speech API. https://github.com/pndurette/gTTS(2018/01/15参照)

[7] Toshiyuki Takezawa, Eiichiro Sumita and Fumiaki Sugaya et al. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. in Proc. of the Third Int. Conf. on Language Resources and Evaluation (LREC). Las, 2002, pp.147-152.

[8] Papineni, Kishore and Roukos et al. BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL ’02, 2002, Philadelphia, Pennsylvania, pp.311-318.

[9] Lidia Mangu, Eric Brill and Andreas Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. CoRR, cs.CL/0010012, 2000.