発話困難者に向けた会話支援システムのためのニューラル機械翻訳を用いた文章推定方式

Sentence Estimation for Conversation Aid System for Speech Disabilities using Neural Machine Translation

渡辺 淳
Jun Watanabe

岡山大学 阿部研究室
Abe Laboratory, Okayama University

概要 発声障害や構音障害によって言語を話すことが困難な方の会話支援を目的とし，タッチタイピングを利用した入力方式による曖昧な入力から話したい文章を推定する方式を検討する．ウェアラブルデバイスからの入力を想定し，得られた入力系列を Attention 機構を用いたニューラル機械翻訳によって翻訳することで，ユーザが発話したい文章を推定する．本報告では，システム自体の評価のため日本語だけでなく英語での評価も行う．

1 はじめに

声を使ったコミュニケーション，すなわち「会話」は，最も手軽かつ伝達速度の速い手段であり，自然言語を使用する人間にとって欠かせないコミュニケーション手段である．「会話」という言葉には手話・ジェスチャーなどによるコミュニケーションを含めるものもあるが，ここでは声を用いたコミュニケーションと同義とする．本研究では，咽頭摘出手術などによる発声機能障害，または舌摘出手術などによる構音機能障害等によって言語を十分に発声できない方 (以下，発話困難者と呼ぶ) の会話を支援するシステムを提案した．

提案するシステムでは，ウェアラブルデバイスによるテキスト入力と，テキスト音声合成 (TTS: Text to Speech) を利用することを想定している．ウェアラブルデバイスについては，両手の十指にスイッチの内蔵された装置を装着し，タッチタイピングの要領で入力を行う．ウェアラブルデバイスを導入する利点として，場所を選ばないシステムの利用が可能になることが挙げられる．筆談等の既存の手段も考えられるが，道具を必要とする手段では，屋外での使用は困難となる可能性がある．また，TTS を用いる場合，スマートフォンなどによってテキスト入力を行う場合も考えられるが，ウェアラブルデバイスを使用することによってアイズフリーの実現が可能となる．つまり，画面を見る必要がなくなり，会話相手の目を見てコミュニケーションをとることができる．このような技術を用いることによって，より健常者どうしの会話に近い日常レベルの会話を目指す．

上記のウェアラブルデバイスから得る入力は 0 から 9 の数字のみであり，TTS への入力とするためにはこの入力された数字を平仮名列，または仮名漢字混じり文に変換する必要がある．数字入力のパターンは，QWERTY 配列キーボードにおけるタッチタイピングを基にしている．手話のようなジェスチャーも伝達速度が早く利点も多いが，多くのジェスチャーを話し手だけでなく聞き手も覚える必要があり，習熟コストが大きい．そのため，最も一般的に普及している QWERTY 配列のタイピング方式を基礎とし，習熟コストの低下を図る．

本研究ではこのウェアラブルデバイスから入力された数値列を，自然言語の文章へ変換する部分を主に扱うものとする．扱う言語については英語と日本語の 2 か国語を扱う．最終的なタスクは日本語での実用化であるが，言語的特徴の観点から英語の方が評価を行いやすいため，システムそのものの評価のために英語での実験も行う．評価実験では，単語正解精度による提案手法の評価と，推定結果例を観察した上での考察を行う．

2 ニューラル機械翻訳の概要

本研究で使用するニューラル機械翻訳（NMT: Neural Machine Translation）モデルはシンボルの系列を入力とし，シンボルの系列を出力とする [1]．タスクによって何をシンボルとして扱うかは異なるが，本研究では入力数値列や英語の単語を指す．実際の処理では単語そのものではなく，単語に割り振られた ID のようなものをシンボルとする．つまり，系列とは単語の系列，すなはち文章を指す．それぞれのシンボルはネットワークの入力時には語彙の one-hot vector で表現される．そのため，学習前に語彙ファイルを用意する必要がある．

ここで入力系列を $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_I)$, 出力系列を $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_J)$ とする．$\mathbf{x}_i$ は入力系列 $\mathbf{X}$ の i 番目の単語を表す．seq2seq モデルの目的は確率 $P(\mathbf{y}_j|\mathbf{Y}_{<j}, \mathbf{X})$ をモデリングすることである．そのモデルは以下の式によって表すことができる．

$$P_\theta(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^{J+1} P_\theta(\mathbf{y}_j|\mathbf{Y}_{<j}, \mathbf{X}) \tag{1}$$

これは言い換えるならば，入力系列と j 番目までの出力系列の単語を基に $\mathbf{y}_j$ の確率を求めるということである．ただ実際には，入力系列 $\mathbf{X}$ は自身から生成された固定長ベクトル $\mathbf{z}$ として入力される．

Encoder 側では，各単語の one-hot vector に embedding を行い，LSTM の入力とする．得られる隠れベクトルを次の単語を処理する LSTM に入力し，これを繰り返す．Decoder 側においても基本的には同様の処理を行うが，Encoder 側との違いは初期隠れベクトルに Encoder 側で最終的に出力された隠れベクトルを用いている点である．これによって，入力系列の情報を保持しつつ処理を行う．最後の出力層では softmax 関数を利用し，確率を算出する．

本研究で使用するモデルの概要図を図 1 に示す．今回の NMT モデルでは，Attention 機構 [3] を用いており，入力系列と出力系列との単語間の関連度合をより強固に学習することが可能となる．1 の赤矢印が Attention 機構の重要な点である．Encoder 側の Hidden vector を全て結合し，Decoder で出力した Hidden vector との内積を求める（和や平均を用いる場

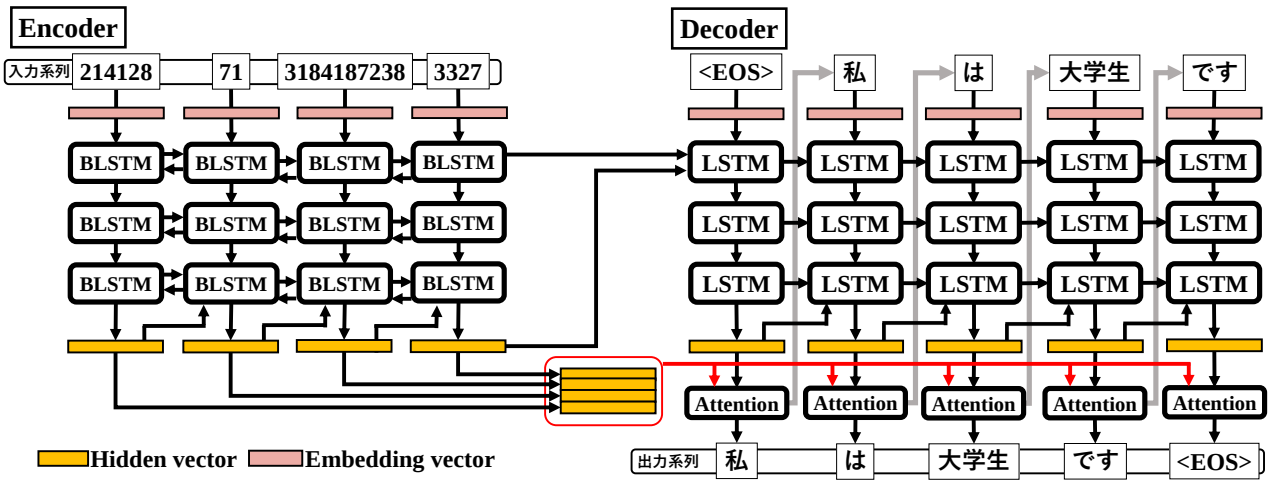

図 1: ＮＭＴモデルの概要図

合もある）．得られた内積を基に context vector を生成し，入力文と出力文の単語のアライメントを学習する．

3　提案手法

提案手法は学習部と翻訳部に分かれており，学習部で作成したモデルを翻訳部で用いる．入力された数値列は，ニューラル機械翻訳と学習済みモデルによって日本語仮名漢字交じり文に変換される．学習済みモデルについては，5 節で後述するが，数値列と日本語文を対訳データとした学習データを学習したものである．

入力は 0 から 9 の十種類の数値からなる数値列であり，以下のように単語区切りの文単位で行われる．

```
214128 / 71 / 3184187238 / 3327
（ 私 / は / 大学生 / です ）
```

それぞれの指には図 2 のように 0〜9 の数値が対応しており，QWERTY 配列のキーボードでタッチタイピングを行う要領で数値を入力する．それぞれの数値は図のように縦の複数個のキーに対応している．例えば，「Hello」という文字列は，「73999」という数値列に対応する．また，単語の区切りは，スペースで示す．すなはち，入力は 5（左の親指）となる．また，文の終わりは，改行で示す．つまり，入力は 6（右の親指）となる．

3.1　変換先の競合問題

提案方式の主な問題点として，入力された数値列に対して変換先の単語が競合する場合があることが挙げられる．これは，アルファベット 26 文字を，0 から 9 の 10 種類の数値で表現するという曖昧性に起因する．つまり，QWERTY キーボードにおける縦列の情報が欠落しているということである．例として，「朝」という単語を意図して「121」と入力した場合を考える．しかしこの場合，「121」の変換先として「泡」という単語の可能性も生じる．このように，入力された数値列の変換先が複数存在するという可能性がある．この問題を解決するためには前後の単語を考慮することが重要であり，ここに NMT を用いる意義がある．

3.2　言語的特徴による問題点

提案手法はあくまで単語を単位として処理を行う．そのため，扱う文章は単語に分かち書きされている必要がある．これは自然言語処理における一般的な問題

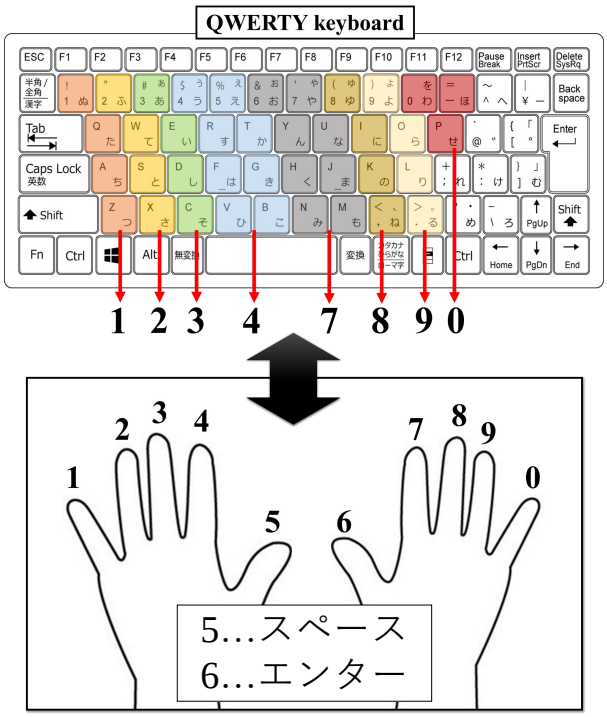

図 2: 各指の数値割り当て

であるが，英語は元々分かち書きされているが，日本語はされていない．さらに，日本語は漢字を含む文章であり，コンピュータに読みを完璧に認識させることも困難である．そのため，日本語の前処理の段階で言語的なミスを含む可能性があるという問題点が存在する．

4　コーパス作成について

学習および検証・テストに用いる対訳データの作成について説明する．本報告では，オープンソースの日英対訳データである JESC(Japanese English Subtitle Corpus)[4] と，EMNLP（Empirical Methods in Natural Language Processing）より wmt17[5] の英語データを利用する．

表 1: 各モデルのサイズ

モデル名	wmt17_2000	wmt17_280	JESC_En	JESC_Jp
学習データ	20M	2.8M	2.8M	2.8M
検証データ	2,000	2,000	2,000	2,000
テストデータ	2,000	2,000	2,000	2,000

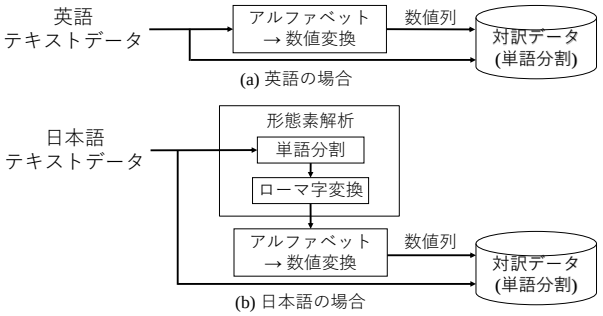

図 3: コーパス作成の概要

4.1 JESC

JESC は日英（英日）翻訳タスクのためのコーパスであり，テレビ番組の字幕データによって構成されている．具体的には，映画やアニメなどの字幕が大半を占めるため，口語的表現の文章を多く含む．

また，JESC はオープンソースな口語表現コーパスとしては最大規模のコーパスである．NMT は長文であるほど翻訳精度が低下する事が報告されており，本タスクにおいてもニュース文などのように長文では動作するとは限らない．その点，JESC は口語的な表現が多く，文の長さもあまり長くないと予想されたため，本タスクの最初の試みとしてこのデータベースを利用した．

次に，データセットの作成手順を示す．

1. JESC から日本語文章を抽出
2. 句読点，記号等の削除
3. MeCab[6]（IPA 辞書）を用いて単語分割
4. MeCab によって読み情報（片仮名列）に変換
5. かなローマ字変換に従い，アルファベットに変換
6. 図 2 に従い，数値に変換
7. 3 と 6 を対訳データとすることでデータセットを作成

英語の場合は，3 と 4 の処理をスキップすることで，対訳データを作成する．

4.2 wmt17

EMNLP では共通のデータセットを用いて，翻訳・要約・質疑応答など様々なタスクにおける言語モデルの性能を評価するカンファレンスがあり，今回，翻訳タスクにて公開されている英語のテキストデータを使用した．wmt17 は新聞記事などのデータを主として構成されており，文語的表現の文章を多く含む．対訳データの作成手順については，JESC の場合と同様である．図 3 にコーパス作成の概要図を示す．

表 2: 学習条件

LSTM ユニット数	1,024
LSTM レイヤ数	3
ミニバッチサイズ	128
最大語数	50
語彙数	50,000
活性化関数	ReLU
損失関数	Softmax Cross Entropy
最適化手法	Adam

5 評価実験

5.1 実験条件

実験に向け，4 つのモデルを作成した．表 1 に，各モデル名とそのサイズを示す．JESC は全体で 280 万ペアのデータセットであるため，wmt17 でも比較用に同規模のモデルを用意した．今回は，学習したモデルを用いてテストデータを翻訳し，その翻訳精度に注目するが，評価する際のテストデータも複数用意した．例えば，wmt17 のモデルを JESC の英語テストデータで評価するなど，文章の特徴が異なる場合の結果にも注目する．

また，表 2 に NMT の学習条件を示す．Encoder では Bi-LSTM を使用し，Decoder では単方向の LSTM を使用する．

5.2 評価尺度

作成したコーパスを元に 4 つのモデルを学習し，評価を行った．評価には以下の単語正解率を一文毎に計算し，それらの平均をとる．

$$\text{単語正解率} = \frac{\text{正解単語数}}{\text{総単語数}} \times 100 \qquad (2)$$

6 結果

6.1 評価尺度による評価

表 3 に，各モデルの単語正解率の平均を示す．いずれの場合においても概ね 9 割前後の高い精度を示しており，提案手法の有効性が示された．

モデルの学習に用いた学習データとテストデータに着目すると，両者が異なるタスクの場合（口語的文章と文語的文章など），正解率が低下するという結果となった．例えば同じ英語であっても，ニュース記事（wmt17）で学習したモデルをＴＶ字幕テキスト（JESC）で評価した場合には，5% 前後の正解率の低下が見られた．この実験からわかるように，翻訳性能はデータセットの特徴に大きく左右される．提案手法は，使用する場面をある程度絞ることでさらに高い正解率を示すことが期待できる．

表 3: 単語正解率の平均

Test data \ Model	wmt17_2000	wmt17_280	JESC_En	JESC_Jp
wmt17	96.00%	95.71%	89.25%	＼
JESC（英）	90.82%	90.34%	96.89%	＼
JESC（日）	＼	＼	＼	92.73%

表 4: 推定結果例（英語）

estimated	I can sometimes see more of the pitch than **they** and they will turn round and give me a thumbs up
answer	I can sometimes see more of the pitch than **them** and they will turn round and give me a thumbs up
estimated	Tragic diana **<unk>** had been in a coma since the crash which caused carnage in the city of **<unk>** eight days ago
answer	Tragic diana **berchenko** had been in a coma since the crash which caused carnage in the city of **kharkiv** eight days ago
estimated	It is fantastic that we will now be able to offer our customers more choice of travel with a world class airline providing **cash** connections to destinations across the world
answer	It is fantastic that we will now be able to offer our customers more choice of travel with a world class airline providing **easy** connections to destinations across the world
estimated	He centered the **<unk>** **<unk>** over her pupil and lowered the flap to seal it in place
answer	He centered the **raindrop inlay** over her pupil and lowered the flap to seal it in place

表 5: 推定結果例（日本語）

estimated	それに 会社 の 決定 に 逆らう 気 も ない
answer	それに 会社 の 決定 に 逆らう 気 も ない
estimated	ベガ と 対決 し た わり に は さほど 服 は 汚れ なかっ た わ ね
answer	ベガ と 対決 し た 割 に は さほど 服 は 汚れ なかっ た わ ね
estimated	無駄 に は し ない わ
answer	無駄 に は し ない さ
estimated	ここ は 山 の 中 です が
answer	ここ は 花 の 中 です が
estimated	映画 会社 は ワーナー ブラザーズ と **<unk>** が 関わり まし た
answer	映画 会社 は ワーナー ブラザーズ と パラマウント が 関わり まし た

6.2 推定結果例からの考察

表 4 に英語の推定結果例を，表 5 に日本語の推定結果例を示す．正解データと異なる出力である単語を赤字で表している．

これらの結果を見ると，出力が異なる場合には大きく 2 つのパターンに分けられる．一つ目は，3.2 節で述べた通り入力される数値列が一致する場合である．このミスは，学習データ中の単語同士の共起回数などに左右される．しかし，日本語の場合には「わり」と「割」のように NMT では別の単語として扱われているが，読みとしては同じ場合もある．さらに，本研究は音声合成によるコミュニケーションを目的としているため，このような場合においては問題視しない．

2 つ目は，<unk> というトークンで表される未知語である．今回，学習データ中で出現頻度の少ない単語は除外し，未知語と推定する実装となっている．しかし，音声合成を行う際に未知語を入力とするのは問題がある．

7 まとめ

今回，NMT による曖昧な入力数値列からの文章推定を行った．日本語の分かち書きされていないという言語的な特徴から，比較のため日本語と英語の両言語で実験を行った．結果としては英語の方が高い正解精度を示したが，日本語においても 92.7% という高い正解精度を示した．

今後の課題としては，未知語等の対処について検討することが挙げられる．実際に音声合成を行う際に，未知語が入力されてしまうと正しく音声が合成できない．そのために，未知語を出力しないようあらかじめ何らかの処理を実装する必要がある．また，実用化に向けて会話内容を想定するなど，学習データを調整することで極力翻訳ミスをしないモデルを学習することも重要である．

参考文献

[1] I. Sutskever, *et al.*, ”Sequence to sequence learning with neural networks” In Advences in Neural Information Processing System (NIPS 2014).

[2] D. Bahdanau et al. “Neural Machine Translation by Jointly Learning to Align and Translate”. In ICLR 2015

[3] M. Luong. et al. “Effective approaches to attention-based neural machine translation”. In EMNLP p.1412-142,1 2015

[4] R. Pryzant et al. ”JESC: Japanese-English Subtitle Corpus”. In Language Resources and Evaluation Conference(LREC) 2018

[5] O. Bojar “Findings of the 2018 Conference on Machine Translation (WMT18)”. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers: 272-303.

[6] 工藤拓 et al. “Conditional Random Fields を用いた日本語形態素解析”. 情報処理学会研究報告. NL, 自然言語処理研究会報告 2004