入力音声に続く文章の予測

恒松 和輝[1,a)]　サクリアニ サクティ[1,2,b)]　中村 哲[1,2,c)]

概要：近年の技術の進歩により，音声認識は人間に近い非常に高い性能を示している．しかし，それらは与えられた音声をただ文字に起こすだけである．人間同士の対話では，発せられた言葉からその後に続く言葉を予測できることがある．本研究では，深層学習を用いてそのようなタスクを実行できるシステムの構築を目指す．本論文では，音声を用いた予測と，音声を用いない文章のみでの予測を行なった．音声を用いた予測では，参照文とは異なるが自然な文章を出力できることが分かった．

1. はじめに

音声認識では，深層学習を用いるのが主流となり，最近は Recurent Neural Network（RNN）のみで音声認識を行う End-to-End の枠組みが研究されている [12]． 最新の音声認識システムでは，Switchboard 会話音声認識のタスクにおいて単語誤り率（WER）5.1%を達成し，IBM が提唱する人間の認識精度と並んだ [11]．人間同士の対話では，相手が最後まで話さなくても，その内容を予測しながら話を聞くことがしばしばある．例えば，図 1 に示すように，一方が「私が食べたのはハンバ…」まで発声したが，「ハンバーガー」という単語を思い出せず，そこで発声が止まってしまった場合を考える．この時，もう一方の聞いている側は，途中まで発声された音声から「ハンバーガー」という単語を予測し，「私が食べたのはハンバーガーです」という内容を理解することができる．しかし，現在の音声認識システムは，入力音声に対して対応する文字列を出力するものであり，上記のような中途半端な音声入力に対しては，その後に続く文字列を予測することはできない．本研究の目的は，このように途中まで発声された音声から，その後に続く内容を予測することである．また，失語症という脳機能障害がある．この失語症のうち，人の言うことを理解したり話をすることに障害は少ないものの，適切な言葉を思い出したり品物の名前を言ったりすることが際だって難しいものを喚語障害（健忘失語）と言う．その思い出せない単語をコンピュータを用いて予測することができれば，喚語障害者がコミュニケーションを行う上で大きな助けとなる．

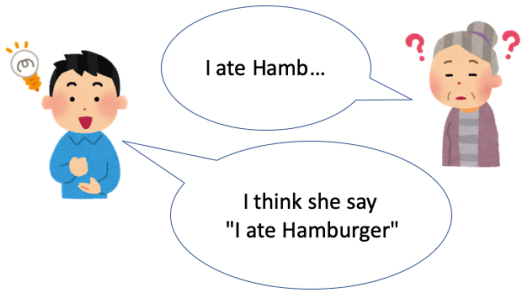

図 1　会話における予測

2. 関連研究

このように，完全な入力を待たずに結果を出力する問題として有名なものに同時通訳がある．同時通訳は一般の翻訳とは異なり，発話に対して出来るだけ遅延なく翻訳を行う必要がある．Jan らは，翻訳元の先頭 n 単語から，完全に訳せる翻訳先の文の長さを，単語の対応関係から割り出し，部分的な入力に対する想定出力のペアデータを作成した．このデータを用いることで，部分的な入力に対する翻訳を実現した [7].

翻訳における予測の例を表 1 に示す．これは英語からスペイン語への翻訳の例で，入力である英語のテキストが部分的になっている場合について示している．部分的な入力に対する出力には 2 パターンあり，上の例では，入力された英語が”I encourage all of”と途中で切れたような入力になっているのに対し，翻訳されたスペイン語の出力結果は足りない部分を予測して補完した形になっている．これ

[1] 奈良先端科学技術大学院大学
NAIST, Takayama-cho, Ikoma, Nara 630–0192, Japan
[2] 理化学研究所 革新知能統合研究センター, RIKEN-AIP
[a)] tsunematsu.kazuki.tj5@is.naist.jp
[b)] ssakti@is.naist.jp
[c)] s-nakamura@is.naist.jp

English: I encourage all of
Spanish: yo animo a todo el mundo.
English: now，I should
Spanish: ahora debera，debera，debera.

表 1　翻訳における予測 [7]

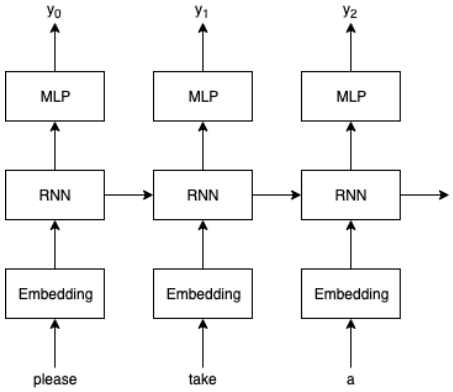

図 2　RNNLM

は予測に成功しているパターンだが，それ以外の場合は，下の例のように最後の単語を繰り返してしまうことが報告されている．

3. 文章のみでの予測

本研究では，音声を用いて予測を行うが，比較のため音声を用いず文章のみでの予測も行なった．モデルには Recurrent Neural Network Language Model（RNNLM） と Bidirectional Encoder Representations from Transformer（BERT）を用いた [3,6].

3.1 RNNLM

RNNLM は図 2 に示すように RNN を用いた言語モデルである．Embedding 層は入力された単語を，分散表現へと変換する．RNN はその分散表現を受け取り，隠れ状態を Multi Layer Perceptron（MLP）と次の RNN へ出力する．こうすることで RNN は過去の情報を保持することができるため，例えば y_2 の出力は，please，take，a の 3 つの入力を考慮したものとなる．RNNLM での予測では，出力の長さを決めず， 文の終わりを示す [EOS] トークンが現れるまで出力を行なっている．

3.2 BERT

BERT は言語表現を事前学習する手法のひとつである．モデルの詳細を図 3 に示す．これまでの言語モデルは，系列中の次の単語を予測するのに，左から右へ単方向の学習を行っていた．一方，BERT は，双方向の Transformer モデル [10] のエンコーダを用いており，転移学習を行うことで質疑応答やタグ付け等の 8 つのタスクにおいて，先行する言語モデルを凌駕する性能を実現した．BERT で利用

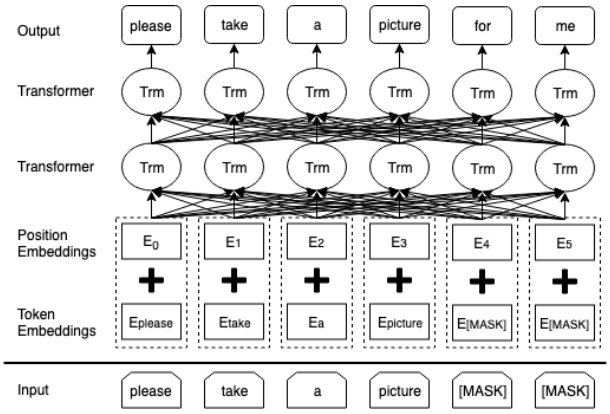

図 3　BERT

する事前学習タスクにはマスクされた言語モデルと後続する文の予測の 2 つがあり，本研究では，このうちマスクされた言語モデルのみを利用する．マスクされた言語モデルは入力系列の一部を [MASK] トークンに置き換え，それを予測するタスクである．本研究では，ランダムに置き換える代わりに，欠落した音声に対応する部分を [MASK] トークンに置き換え，予測を行う．その際に，予測するトークンの数を事前に知ることができないため，[MASK] トークンの数を，データ中で最大の文の長さと同じになるよう設定し，目的のトークンの数が [MASK] トークンの数よりも小さければ，[PAD] トークンで埋めている．また，本研究では，分類タスクや後続する文の予測は行わないため，分類埋め込みのための [CLS] トークンや，文を分ける [SEP] トークンは用いず，入力毎に文を分けている．

モデルのハイパーパラメータや重みの初期値は，公開されている BERT-Base モデル[*1]を用いた．このモデルは，12 層，768 次元の隠れ層，12 個の Self-Attention ヘッドで構成されている．

4. 提案手法

本研究では，音声を用いた予測を行う． 予測には，図 4 に示すエンコーダーデコーダ（Encoder-Decoder）モデルを用いる [1,2,8,9]．エンコーダーデコーダモデルは，エンコーダとデコーダの 2 つのネットワークにより構成されている．エンコーダ側では，入力音響特徴量 $\mathbf{x} = [x_1, x_2, ..., x_N]$ をフレーム毎に数値ベクトル（分散表現）$\mathbf{h} = [h_1, h_2, ..., h_n]$ にエンコーディングする. 本研究では，エンコーダに多層 RNN ネットワークを採用しており，計算コスト削減のために，2 層以降の処理では前の層の出力 $\mathbf{h}$ のうち偶数番目だけを抽出し入力とする処理を繰り返した．デコーダ側では，このエンコードされた分散表現系列 $\mathbf{h}$ を入力として出力を順次予測する．その時に，音声認識では最初の方の文字は音声の最初の方の情報を用いるのが有用であるため，各分散表現系列 h_t に対して重み a_t を付ける．これが

[*1] https://github.com/google-research/bert

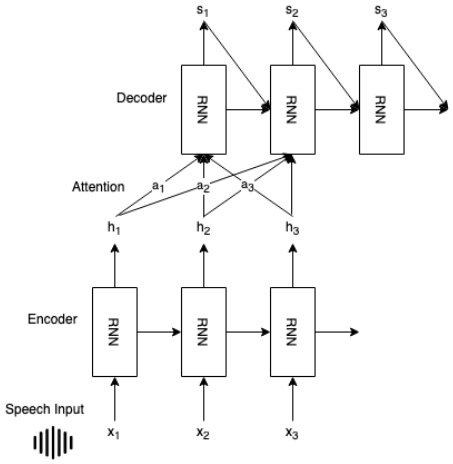

図 4　エンコーダーデコーダモデル

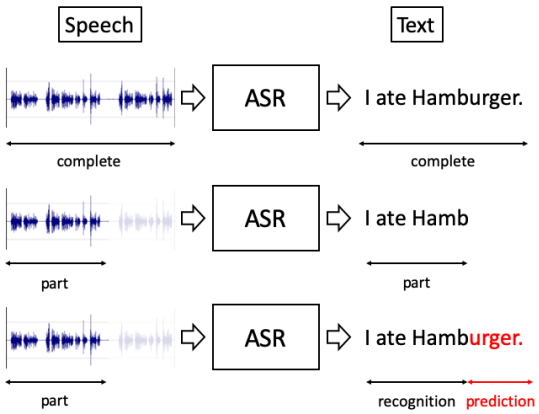

図 5　欠落した音声からの文字予測

Attention である．エンコーダは音響モデル，デコーダは言語モデル，Attention は入力と出力の関係を表している．本研究では，Attention 機構に MLP を使用している．

$$a_t = W_s * (tanh(W_c * (h_t + h^{tgt}))) \qquad (1)$$

ここで W_s と W_c は学習可能なパラメーターであり，h^{tgt} はデコーダ側の分散表現を表している．通常のエンコーダ-デコーダモデルでは，Attention は入力されたエンコーダ系列全体の中から最適なベクトルをデコーダの入力をもとに見つける．しかし，図 5 に示すように，本研究では音声の入力は一部しか与えられないため，デコーダは，音声が欠落していて入力のない部分に対しても何らかの出力を行うよう Teacher Forcing で学習させてある．また，この欠落した音声に対し，モデルを学習することによって，各モデルはどれだけの長さの文章を出力するべきかを学習することができる．本来ならば，音声がどれだけ欠落するかは未知であるため，文章をどれだけ生成するかも未知の問題であるが，本研究では，音声の欠落部分に対する出力文の長さではなく，内容に着目しているためこのような設計にした．

Corpus	Number of sentences		
	Train	Val	Test
BTEC	157448	4870	510

表 2　BTEC コーパスのサイズ

Input	Model	InputLength	WER(%)
Text Input	RNNLM (Baseline)	75%	22.34
		50%	99.79
		25%	106.40
	BERT	75%	8.51
		50%	21.37
		25%	48.60
Speech input	ASR Predict	75%	4.46
		50%	20.98
		25%	42.12
	ASR(Topline)	100%	2.03

表 3　各モデルの WER

5. 実験

本研究では，欠落した音声を入力したとき，モデルは欠落部を生成する．その生成された部分も含めた文全体の自然さと認識率を評価した．また，文章のみでの予測も同様に行なった．文章のみの予測では，参照文のうち，音声が存在する部分の単語を入力とした．

5.1 データセット

予測を行うデータには Basic Travel Expression Corpus（BTEC）[4,5] のテキストから，Google 音声合成を用いて音声を作成し，利用した．表 2 に用いたデータ数の詳細を示す．

5.2 実験設定

通常の音声の全体の長さを 100%としたとき，それを後ろから 25%ずつ 3 段階に渡って削除したものを用意する．用意した音声は，短いものから 25%，50%，75%となり，それぞれについて学習と評価を行う．各削除レートごとにモデルを分けて学習するため，モデルは学習を通して，入力音声にどれだけの欠落があるかを獲得することができる．この情報を用いて，モデルは入力に対して一定の過剰生成を行う．

5.3 単語誤り率による予測結果の比較

欠落した 3 種類の音声を，通常の音声認識で認識した結果と，それぞれ学習を行なった場合の予測結果，また音声を使用せず文章のみでの予測結果の単語誤り率（WER）を表 3 に示す．

5.4 主観評価による出力文の自然さ比較

参照文，音声認識結果，欠落した 3 種類の音声の予測結

Input	Model	InputLength	Naturalness
Text input	RNNLM (Baseline)	75%	2.76
		50%	2.29
		25%	2.49
	BERT	75%	3.87
		50%	3.04
		25%	1.43
Speech input	ASR Predict	75%	4.44
		50%	3.92
		25%	4.32
	ASR(Topline)	100%	4.76
	Reference		4.78

表 4 各モデルの出力文の自然さ

果，音声を使用しない予測結果のそれぞれについて，各 5 文ずつ，出力文の自然さを 5 段階で評価した平均の値を表 4 に示す．評価は TOEIC スコア 730 点以上を持つ 12 人に行なってもらった．

5.5 考察

出力文の自然さ比較において，音声予測結果が 50%よりも 25%の方が高くなっているのは，入力音声が長いほど，後に続く文章が参照文と異なる内容だったときに文全体として不自然になってしまうことに起因すると考えられる．また，いずれの場合も単語誤り率が高く，参照文とは異なる出力をしていることが分かる．しかし，音声予測結果では文全体の自然さは高く，参照文とは異なるものの，自然な文章を出力していると言える．

6. おわりに

本論文では，欠落した入力音声から，それに続く文章の予測を行った．また，比較のため音声を用いない文章のみでの予測も行なった．音声を用いた予測では，参照文とは異なるが自然な文章を出力できることが分かった．ただし，本論文の実験では音声を用いた場合と用いない場合で使用しているモデルが異なるため，適切な比較とは言えず，同じモデルを用いた比較を行う必要がある．

7. 謝辞

本研究は科研費 JP17H06101，JP17K00237 の助成を受けております．

参考文献

[1] Bahdanau, D., Cho, K. and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Bengio, Y. and LeCun, Y., eds.), (online), available from ⟨http://arxiv.org/abs/1409.0473⟩ (2015).

[2] Chan, W., Jaitly, N., Le, Q. V. and Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pp. 4960–4964 (online), DOI: 10.1109/ICASSP.2016.7472621 (2016).

[3] Devlin, J., Chang, M., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186 (online), available from ⟨https://aclweb.org/anthology/papers/N/N19/N19-1423/⟩ (2019).

[4] Kikui, G., Sumita, E., Takezawa, T. and Yamamoto, S.: Creating corpora for speech-to-speech translation, *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*, (online), available from ⟨http://www.isca-speech.org/archive/eurospeech_2003/e03_0381.html⟩ (2003).

[5] Kikui, G., Yamamoto, S., Takezawa, T. and Sumita, E.: Comparative study on corpora for speech translation, *IEEE Trans. Audio, Speech & Language Processing*, Vol. 14, No. 5, pp. 1674–1682 (online), DOI: 10.1109/TASL.2006.878262 (2006).

[6] Merity, S., Keskar, N. S. and Socher, R.: Regularizing and Optimizing LSTM Language Models, *CoRR*, Vol. abs/1708.02182 (online), available from ⟨http://arxiv.org/abs/1708.02182⟩ (2017).

[7] Niehues, J., Pham, N., Ha, T., Sperber, M. and Waibel, A.: Low-Latency Neural Speech Translation, *CoRR*, Vol. abs/1808.00491 (online), available from ⟨http://arxiv.org/abs/1808.00491⟩ (2018).

[8] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to Sequence Learning with Neural Networks, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 3104–3112 (online), available from ⟨http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks⟩ (2014).

[9] Tjandra, A., Sakti, S. and Nakamura, S.: Listening while speaking: Speech chain by deep learning, *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, Okinawa, Japan, December 16-20, 2017*, pp. 301–308 (online), DOI: 10.1109/ASRU.2017.8268950 (2017).

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need, *Advances in neural information processing systems*, pp. 5998–6008 (2017).

[11] Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X. and Stolcke, A.: The Microsoft 2017 Conversational Speech Recognition System, *CoRR*, Vol. abs/1708.06073 (online), available from ⟨http://arxiv.org/abs/1708.06073⟩ (2017).

[12] 河原達也：音声認識技術の変遷と最先端 (2018).