カリキュラムラーニングを用いた音声翻訳の学習戦略の提案

叶 高朋[1]　サクティ サクリアニ[1,2]　中村 哲[1,2]
奈良先端科学技術大学院大学 情報科学研究科[1]
理化学研究所　革新知能統合研究センター[2]
{kano.takatomo.km0, ssakti, s-nakamura}@is.naist.jp

1　はじめに

近年, 国際化により多くの外国人が日本へ訪れるようになり, また, 東京オリンピックの開催に伴い多言語で円滑にコミュニケーションを取る必要性が高まっている. 英語は有用な言語であるが, 各人の発音の差異・英語の習熟度により必ずしも機能しない場合がある. この問題を解決する技術として, お互いの母国語同士の翻訳を可能とする自動音声翻訳技術がある. 従来の音声翻訳は音声認識・機械翻訳・音声合成から構成され, テキストを媒介に受け渡すため, 機械翻訳が音声認識誤りに影響を受ける問題がある [12]. 近年, Duong ら, Alexandre らが深層学習 [11] を用いて, 入力音声から直接翻訳するモデルを, 英語・スペイン語/英語・フランス語について提案している [6, 3]. これらの研究では, 語の並べ替えが限定的で翻訳が容易な言語対を扱っているため, 問題の難しさとして一般的な音声認識の問題変わらない場合がある. 本研究では, 語順の違う差異の大きな日本語・英語の翻訳を扱うため, より複雑な問題を効果的に学習する構造的カリキュラム学習法を提案する. 従来のカリキュラム学習は, 簡単なデータから学習を始め複雑なデータを加えていく学習法で, 複雑な問題の学習に有効な戦略である [2]. ここでは, 問題・モデル構造徐々に拡張しながら学習する方法を提案する.

2　音声翻訳について

注意型シーケンシャルモデルを下記のように構築した [1]. 長さ N の入力系列 $x = [x_1, x_2, ..., x_N]$ に対し, 長さ T の出力系列 $y = [y_1, y_2, ..., y_T]$ とその条件付き確率 $p(y|x)$ は, 下記の通り表される.

$$p(y_t|y_1, y_2, ..., y_{t-1}, x) = softmax(h_t^{dec}). \quad (1)$$

W_y はデコーダの隠れ層から, 目的言語の語彙数次元への線形写像の重みである. デコーダーの隠れベクトル h_t^{dec} は, t 番目の出力を生成するためのコンテキスト情報 c_t と重み W_c を用いて下記のように表される.

$$h_t^{dec} = tanh(W_c[c_t; h_t^{dec}]). \quad (2)$$

ここで, c_t はでアテンションモジュールにおいて以下のように得られる.

$$c_t = \sum_{n=1}^{N} at(n) h_n^{enc} \quad (3)$$

$$a_t(n) = align(h_n^{enc}, h_t^{dec}) = softmax(dot(h_n^{enc}, h_t^{dec})). \quad (4)$$

h^{enc} はエンコーダの出力系列であり, 本研究では双方向 long short-term memory (bi-LSTM) を用い, デコーダは単方向 LSTM を用いた. アテンションモジュールでは, デコーダの隠れ状態に基づき出力に有用なエンコーダーの情報を推定している. $align(h_n^{enc}, h_t^{dec})$ の計算方法にはいくつか種類があるが, ここではエンコーダーの隠れ層の系列とデコーダの隠れ層の内積を用いた [11].

3　提案手法

注意型シーケンシャルモデルの学習は, 一般的なニューラルネットワークモデルの学習と比べ, エンコーダ, デコーダ, アテンションの 3 つのモジュールを同時に最適化する必要があるため難しいとされている [4]. また, 音声翻訳では, 音声認識で扱われる長い入力系列の境界を推定し, 出力単語に紐付ける問題 [5] と, 機械翻訳で扱われる入力単語と出力単語の対応関係発見し, 変換・並べ替えルールを学習する問題 [1] を同時に解く必要がある. 本研究では, 入力音声から直接対訳文を出力する注意型シーケンシャルモデルを学習す

る際に, 従来のカリキュラム学習ではなく, 音声認識, 機械翻訳といった比較的簡単なタスクから学習し, 構造を組み替えながら最終的に入力音声から出力文を直接翻訳するモデルを学習する構造的カリキュラム学習を提案する図 1. に, 注意型シーケンシャルモデルに対しどのように段階的に構造的カリキュラム学習を進めていくかを示す.

- FastTrack
  - フェーズ 1: 音声認識の学習を行う.
  - フェーズ 2: 音声認識モデルのデコーダ部分を, 機械翻訳デコーダで置き換え音声翻訳の学習を行う.
- SlowTrack
  - フェーズ 1: 音声認識と機械翻訳の学習を行う.
  - フェーズ 2: 音声認識モデルのトランスコーダに置き換え, 機械翻訳のエンコーダ出力を教師としてトランスコーダの学習を行う.
  - フェーズ 3: トランスコーダに機械翻訳モデルのアテンションとデコーダを結合し, 音声翻訳の学習を行う.

4 実験設定

実験は, Basic Travel Expression Corpus(BTEC)[7, 8] 英日対訳文のうち, 学習に 4,5000 発話, テストに 500 発話用いた. 入力音声は, Google 音声合成システムを利用し BTEC コーパスから合成して生成した. この音声に対し, 窓幅 25 m s シフト幅 10 m s の解像度で, 23 次元 FilterBank 特徴量を Kaldi を用いて抽出し, 平均 0 分散 1 となるように正規化した後に学習とテストに用いた. また, 音声認識, 機械翻訳, 音声翻訳について, 積層数 2 の LSTM を利用し, LSTM の隠れ層のユニット数は 512, 原言語の語彙数は 27,293 語, 目的言語の語彙数は 33,155 語, 単語のエンベットサイズは 128 に設定した. また, 最適化手法として Adam を用いている [9]. 注意型シーケンシャルモデルを用いて音声認識・機械翻訳・音声翻訳システムを構築し, 提案するカリキュラム学習を適応して学習効果と翻訳精度を計測した.

- Baseline MT:テキストベースの機械翻訳機
- Baseline ASR+MT:テキストレベルで結合した音声翻訳機
- Direct ST Enc-Dec:注意型シーケンシャルモデルを用いた直接音声翻訳機
- Fast Track: 直接音声翻訳機に対して fast track の学習を適応したモデル
- Slow Track: 直接音声翻訳機に対して slow track の学習を適応したモデル

単一話者の合成音声に対する音声認識誤り率は 9.4%であり, 翻訳精度は BLEU+1 を用いて計測した [10]. 最初に, 提案するカリキュラム学習の効果について, 各学習エポックにおける softmax cross-entropy の値を図 2 に示す. 通常の学習方法で直接音声翻訳機を学習した際は, 最も損失の減少がなかった. 一方, 提案するカリキュラム学習を適応すると, Fast Track においては, 同様のモデル設計を用いているにもかかわらず損失が減少した. また, Slow Track では, テキスト機械翻訳を超える損失減少を達成した. 次に, 各モデルについて翻訳精度を計測した結果を図 3 に示す. 実験結果より, 通常の学習法で学習した日英の直接翻訳モデルでは翻訳が困難なことがわかる. 直接音声翻訳機は過度に出力言語の系列情報に適応しており, 入力音声の情報を考慮していない傾向があった. また, 結果より提案した Fast Track では明らかに翻訳精度が向上していること, Slow Track が最も高い翻訳精度達成したことから提案した学習法が期待した効果をあげたと言える. Slow Track は音声認識のエンコーダとアテンション, トランスコーダ, 機械翻訳アテンションとデコーダで構築されており, 機械翻訳の観点から見るとトランスコーダはよりノイズを含む入力を与えるため, De-noising auto encoder のような機能を果たしテストデータに対して頑健になったと考えられる. 一般的な機械翻訳では, 単語の入力として One-hot-vector をもちいる. これをエンベットした際に, 意味/用法の近い単語は特徴空間上近い場所にマッピングする. この時, データによっては異なる二単語が十分に近い場合デコーダ側でアテンションを取る際に間違いを起こすことがある. こういった減少に対し, 音響情報を入力に付与することによって, これらエラーを回避し正しく翻訳できるようになることが確認されている. [13] また従来, 単語という離散的な入力から, 音声認識の学習により写像された連続的な意味空間上のベクトルに

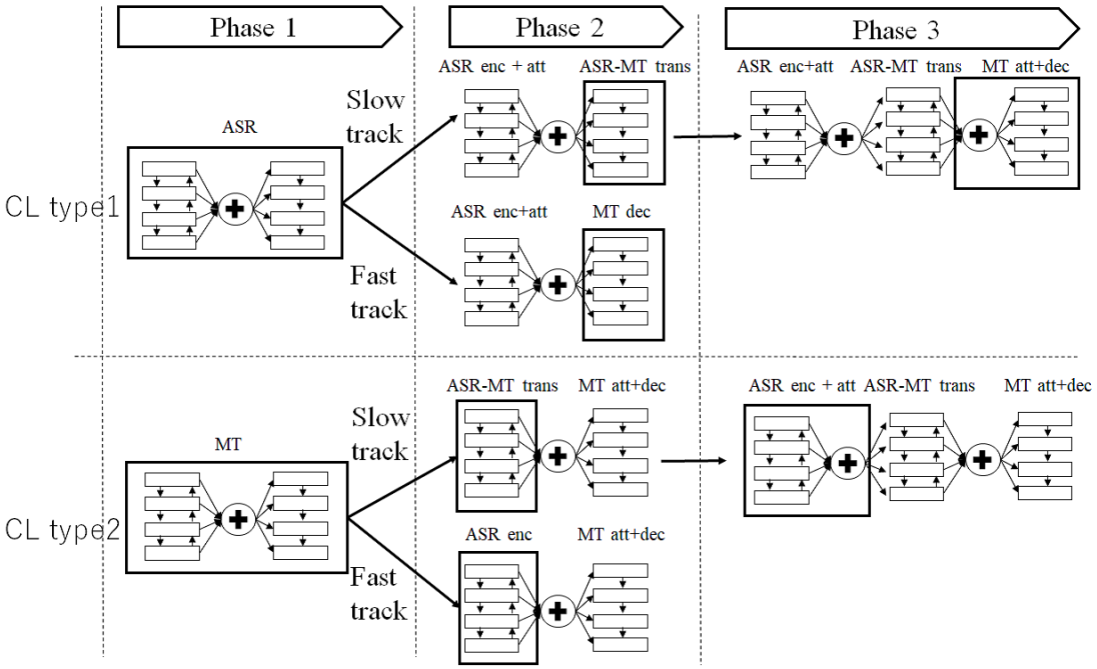

図 1: 提案手法の概要

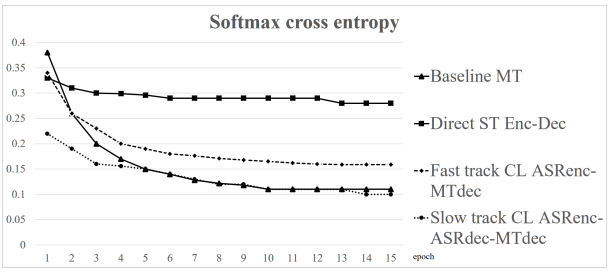

図 2: 各手法ごとの学習の進み方

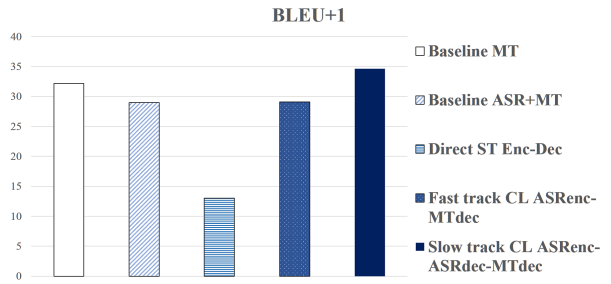

図 3: 各手法ごとの BLEU+1 スコア

なったため, より効率的な学習が可能になったと考えられる.

5 結論

本研究では, 英日の直接音声翻訳を実現した. 提案した直接音声翻訳では原言語のテキストを推定しないため従来の音声翻訳に比べ, 音声認識誤りに影響を受けない. また, 提案した構造的カリキュラム学習は複数の簡単な問題に分解できる, 複雑な問題に対する学習に効果があり, 最終的にテキスト機械翻訳と同等もしくは僅かに良い性能を示した. 一方で, 本研究は合成された単一話者の音声を用いているため, 今後は一般音声への適応や様々な言語対の翻訳精度の確認, 音声合成部の拡張や非言語情報の取扱などを討していく.

6 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] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, Vol. abs/1409.0473, , 2014.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pp. 41–48, 2009.

[3] Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *CoRR*, Vol. abs/1612.01744, , 2016.

[4] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pp. 4960–4964, 2016.

[5] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 577–585, 2015.

[6] Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. An attentional model for speech translation without transcription. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 949–959, 2016.

[7] Gen-ichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. Creating corpora for speech-to-speech translation. In *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*, 2003.

[8] Gen-ichiro Kikui, Seiichi Yamamoto, Toshiyuki Takezawa, and Eiichiro Sumita. Comparative study on corpora for speech translation. *IEEE Trans. Audio, Speech & Language Processing*, Vol. 14, No. 5, pp. 1674–1682, 2006.

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, Vol. abs/1412.6980, , 2014.

[10] Chin-Yew Lin and Franz Josef Och. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland*, 2004.

[11] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *CoRR*, Vol. abs/1508.04025, , 2015.

[12] Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, Jin-Song Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. The ATR multilingual speech-to-speech translation system. *IEEE Trans. Audio, Speech & Language Processing*, Vol. 14, No. 2, pp. 365–376, 2006.

[13] Kaho Osamura, Takatomo Kano, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. Using spoken word posterior features in neural machine translation. In *15th International Workshop on Spoken Language Translation*, pp. 181–188, 2018.