Vol.2021-MUS-131 No.17
Vol.2021-SLP-137 No.17
2021/6/18

End-to-end 音声認識モデルにおける暗黙的言語情報の置換法

森 大輝[1,a)]　太田 健吾[2]　西村 良太[3]　小川 厚徳[4]　北岡 教英[1]

概要：近年，End-to-end 音声認識が従来の DNN-HMM 音声認識と比べ，高速かつ簡潔であることから注目されている．さらに大量のテキストデータによって学習された言語モデルを併用することで，認識精度が向上すると報告されている．本稿では，音声認識モデルと言語モデルの一般的な統合方法とされる Shallow Fusion を応用した新しい言語モデルの統合方法である Language Model Replacement を提案する．提案法では，事前学習済み音声認識モデルと事前学習済み言語モデルを用いる．提案法ではベイズ則に基づき，音声認識モデルに暗黙的に含まれる言語情報を差し替えることが可能となっている．我々の実験では，学術講演音声データを使用して学習された音声認識モデル内部の言語情報を，模擬講演テキストデータで学習した言語モデルによって差し替えた．模擬講演ドメインにおける提案法の CER は Shallow Fusion での認識精度と比較して，1.3 ポイント上回った．

キーワード：End-to-end 音声認識, 言語モデル, 暗黙的言語情報

Language Model replacement method for end-to-end speech recognition which excludes implicit linguistic information

Daiki Mori[1,a)]　Kengo Ohta[2]　Ryota Nishimura[3]　Atsunori Ogawa[4]　Norihide Kitaoka[1]

Abstract: Recently, end-to-end speech recognition has attracted much attention because it is faster and more concise than conventional DNN-HMM speech recognition. It has also been reported that recognition performance is improved by employing a language model trained with a large amount of text data. Based on these observations, we propose a new language model integration method which we call Language Model Replacement. In our proposed method, we use a pre-trained speech recognition model and a pre-trained language model. In contrast to the Shallow Fusion method, our proposed method can replace the linguistic information implied in the ASR model with independently trained model based on Bayes' rule.
In our experiments, the ASR linguistic information implicitly trained using the Japanese language Academic Presentation Speech corpus is replaced with the language model trained using the Japanese language Simulated Public Speech corpus. We then compare ASR performance for Japanese speech recognition tasks using the Character Error Rate (CER). Our proposed Language Model Replacement method achieved 1.3 percent lower CER in comparison to the Shallow Fusion method.

Keywords: End-to-end Speech Recognition, Language Modeling, Implicit linguistic Information

1. はじめに

従来の DNN-HMM ASR システムは，音響，辞書，言語モデルなどの様々なモジュールから構成されており，非常に複雑な構造であった [1,2]．一方で，End-to-end ASR

[1] 豊橋技術科学大学
[2] 阿南工業高等専門学校
[3] 徳島大学
[4] 日本電信電話株式会社
[a)] mori.daiki.ax@tut.jp

モデルは，単一のニューラルネットワークを用いて，音響特徴量系列を記号ラベル系列（文字や単語など）に直接変換するモデルを学習することで，シンプルかつ高速な認識を可能とする．特に，音声認識や機械翻訳などの自然言語処理タスクにおいては，Seq2Seq モデル [3, 4] が注目されており，その性能は著しく向上している．Seq2Seq モデルは，十分な量の学習データがあれば，高精度なモデルを作成することが可能だが，大量のテキストデータで学習した言語モデルを併用することで，豊富な言語情報を活用でき，さらに高い性能を得ることができる．また言語モデルはテキストデータのみで学習可能であるため，対象となるドメインの学習データを大量に準備することは比較的容易である．ASR モデルと言語モデルの標準的な統合方法として，Shallow Fusion [5] と呼ばれる手法がある．この手法は，推論時において，ASR モデルの出力確率と言語モデルの出力確率を，対数領域で加算することで，各モデルを統合する手法である．他にも，事前学習済み ASR モデルと事前学習済み言語モデルのそれぞれの隠れ状態を入力とする DNN を学習する Deep Fusion [6] と呼ばれる手法も提案されている．さらに，Deep Fusion の改良版として，ASR モデル学習時に事前学習済み言語モデルを使用する Cold Fusion [7] も提案されている．このように，ASR モデルと言語モデルの両方を利用する様々なアプローチが存在する．上述したいずれの言語モデル統合手法も Seq2Seq モデルの性能を向上させることは可能となる．しかし，それぞれの統合手法にはいくつかの問題がある．はじめに，Shallow Fusion の場合，ASR モデルのモデルパラメータは学習データに対して正確な推論結果を出力するように推定されるため，ASR モデルには “暗黙の言語情報” が含まれていることになる．そのため，学習データに含まれる言語情報が ASR モデルのモデルパラメータに反映される．従って，Shallow Fusion は学習データに含まれる言語情報に依存する ASR モデルの出力確率に，対象となるドメインの言語モデルの出力確率を対数領域で加算していることになる．次に，Deep Fusion や Cold Fusion では，言語モデルを統合する際に再学習が必要となる．Deep Fusion は，事前学習済み ASR モデルと事前学習済み言語モデルの隠れ状態を入力とした DNN を学習する必要がある．そのため，言語モデルを差し替えたい場合は，DNN のモデルパラメータを再推定する必要がある．同様の理由で，Cold Fusion は，ASR モデルは事前学習済み言語モデルを使って学習されるため，言語モデルを交換する際には ASR モデルの再学習が必要になる．

本稿では，これら 2 つの問題を克服するための言語モデル統合手法として，Language Model Replacement を提案する．提案法は，ベイズ則に基づいて “暗黙の言語情報” が除去された ASR モデルの出力確率と対象ドメインの言語モデルの出力確率を対数領域で加算する手法である．提案法は，学習データを準備することが困難なドメインに対して，そのドメインのテキストデータのみを準備することで，音声認識器を作成することを可能とする．また，Shallow Fusion と同様に，ASR モデルと言語モデルを推論時にのみ統合する手法であるため，言語モデルを差し替える際の再学習は不要となる．

2. 関連研究

2.1 ASR モデルと言語モデルの統合

ASR モデルと言語モデルを統合するための標準的な手法は Shallow Fusion [5] と呼ばれている．この手法は ASR モデルと言語モデルの出力確率を対数領域で加算するもので，以下の式 (1) で定式化される．

$$y = \underset{y}{\operatorname{argmax}}\{\log P_{ASR}(y|x) + \lambda \log P_{LM}(y)\} \quad (1)$$

ここで，$\log P_{ASR}(y|x)$ は，音響特徴量系列 x が入力されたときに記号ラベル系列 y を推定する ASR モデルの出力確率，$\log P_{LM}(y)$ は，言語モデルの出力確率である．Shallow Fusion では，言語モデルは推論時にのみ使用され，言語モデルと ASR モデルは事前に独立して学習される．

また，ASR モデルと言語モデルを統合する別の方法として，Deep Fusion [6] と呼ばれる手法がある．Deep Fusion を定式化したものを以下の式 (2a) から式 (2c) に示す．

$$g_t = \sigma(v^\top s_t^{LM} + b) \quad (2a)$$

$$s_t^{DF} = [s_t; g_t s_t^{LM}] \quad (2b)$$

$$y_t = \mathrm{softmax}(\mathrm{DNN}(s_t^{DF})) \quad (2c)$$

ここで，$[s_t; g_t s_t^{LM}]$ はベクトル s_t とベクトル $g_t s_t^{LM}$ を連結したものである．s_t，s_t^{LM} および s_t^{DF} はそれぞれ事前学習済み ASR モデル，事前学習済み言語モデル，Deep Fusion モデルの隠れ状態を表す．スカラー g_t は s_t と重みパラメータ v および b を用いて学習されたゲート値である．式 (2c) において，DNN は任意の数の層を持つディープニューラルネットワークである．ASR モデルと言語モデルは，事前に学習された各モデルの隠れ状態を入力とする DNN を学習することで統合される．

また，Deep Fusion の改良版として Cold Fusion [7] と呼ばれる手法も提案されている．Cold Fusion において，ASR モデルは事前学習済み言語モデルの言語情報を参照しながら学習される．Cold Fusion は，以下の式 (3a) から式 (3d) で定式化される．

$$h_t^{LM} = \mathrm{DNN}(l_t^{LM}) \quad (3a)$$

$$g_t = \sigma(W[s_t; h_t^{LM}] + b) \quad (3b)$$

$$s_t^{CF} = [s_t; g_t \circ h_t^{LM}] \quad (3c)$$

$$y_t = \mathrm{softmax}(\mathrm{DNN}(s_t^{CF})) \quad (3d)$$

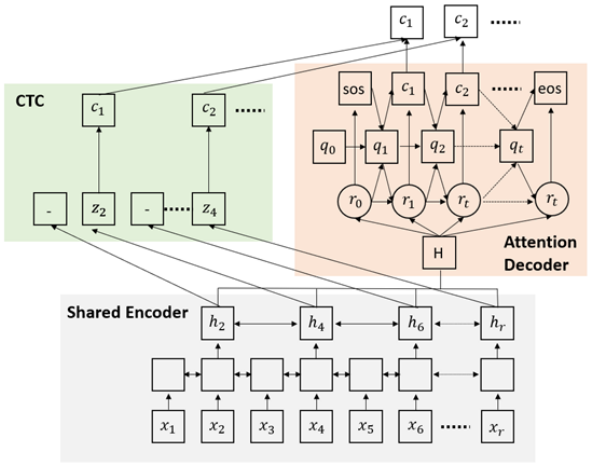

図 1 Hybrid CTC/Attention Architecture

ここで，l_t^{LM} は言語モデルの logit 値，s_t は ASR モデルの状態を表している．ゲート値 g_t は，h_t^{LM}，ASR モデルの状態 s_t および重みパラメータ W，b を用いて学習される．s_t^{CF} は s_t，g_t および l_t^{LM} のアダマール積によって得られるベクトルの連結である．したがって，ASR モデルの状態 s_t と言語モデルの logit 値を入力とした DNN の出力 l_t^{LM} を連結し，それぞれの情報を統合している．また，ゲートアルゴリズムとして fine-grained (FG) gating mechanism [8] を使用することで，Cold Fusion の性能が向上することが報告されている．

前述のように，ASR モデルと言語モデルを統合する手法がいくつか提案されており，これらにより音声認識性能が向上することが分かっている．

2.2 ESPnet

ESPnet は，End-to-end モデルに特化したオープンソースの音声処理ツールキットである [9]．本実験では，ESPnet が提供する RNN ASR モデルを使用した．このモデルは Hybrid CTC/Attention Architecture [10,11] を採用している．図 1 に Hybrid CTC/Attention Architecture のモデル図を示す．

入力された音響特徴量系列は VGGnet により整形される．次に，エンコーダとして使用される 6 層の BLSTM (Bidirectional LSTM) 層によって中間表現 H に変換される．Seq2Seq デコーダは，1 層の LSTM 層および 1 層の Linear 層で構成されている．また，CTC デコーダは 1 層の Linear 層のみで構成される．

ESPnet は，言語モデルの学習にも利用できる．今回の実験では，文字ベースの RNN 言語モデルを使用した [12]．このモデルは，Embedding 層，2 層の LSTM 層，1 層の Linear 層で構成されている．ESPnet では，デコードの際に ASR モデルと言語モデルの Shallow Fusion を行うことができる．

2.3 HMM による音声認識

GMM-HMM および DNN-HMM のような HMM ベースの音声認識システムにおいて，$P(x|y)$ は単語および文字列 y が与えられたときの音声特徴量 x の尤度を表している．y の事前確率が得られれば，ベイズ則を用いて式 (4) のように，これらを組み合わせることができる [13, 14].

$$P(y|x) \propto P(x|y) \cdot P(y). \tag{4}$$

また $P(x|y)$ と $P(y)$ の確率的誤差を正規化するために，式 (5) のように対数領域の重み付け係数 λ を使用することが多い．

$$\log P(y|x) \propto \log P(x|y) + \lambda \log P(y). \tag{5}$$

ここで，$P(x|y)$ と $P(y)$ は，それぞれ音響モデルと言語モデルによって出力され，これらのモデルは独立して学習される．よって HMM ベースの統計的音声認識システムは，モジュール化されている．

2.4 先行研究

本稿では，式 (5) に基づいて，ベイズ則を用いて Shallow Fusion 法を理論的に改善する．我々の研究と並行して，McDermott ら [15] は RNN-T デコーダを用いて，我々の提案法と類似した手法を提案している．RNN-T では，過去の認識結果および現在の音響特徴量から文字列に直接変換される．一方本稿では，同様の考え方を Seq2seq モデルに適用する手法を提案する。実際に、ビームサーチを用いた推論が可能であることが確認している．従って，本手法は，理論的に，Transformer [17] ベースのモデルあるいは Conforemer モデル [18] などにも有効であると考えられる．本研究では漢字かな交じりなど文字システムが比較的複雑とされる日本語データベースを使用して実験を行った．また，[16] では，ベイズ則を用いてエンコーダ・デコーダモデルを定式化してるが，スケーリングおよび正規化係数は考慮していない．本研究では LSTM ベースのエンコーダ-デコーダモデルにおいて，スケーリングおよび正規化係数を考慮し，実験を行った．

3. 提案手法

我々が提案する Language Model Replacement は，Shallow Fusion と類似しているが重要な違いがある．それは我々の提案法は “暗黙の言語情報” を考慮するということである．ASR タスクにおける推論の際，我々は以下の式 (6) ように出力系列 $\hat{y}$ を推論しようとする．

$$\hat{y} = \underset{y}{\operatorname{argmax}} \{\log P_{souce}(y|x)\} \tag{6}$$

ここで，$\log P_{source}(y|x)$ は音響特徴量系列 x が入力されたときに，記号系列 y が出力される確率を表している．“source” とは，ASR モデルの学習に用いる音声データを

録音したドメインのことである．ASR モデルの対数出力確率 $\log P_{souce}(y|x)$ は，ベイズ則を用いて以下の式 (7) のように展開できる．

$$\begin{aligned}\log P_{source}(y|x) &= \log P_{source}(x|y) \\ &\quad + \log P_{source}(y) - \log P_{source}(x) \\ &\propto \log P_{source}(x|y) + \log P_{source}(y)\end{aligned} \tag{7}$$

式 (7) の右辺に注目すると，ASR モデルには音響的な情報の項 $\log P_{source}(x|y)$ と言語的な情報の項 $\log P_{source}(y)$ が含まれていることが分かる．ASR モデルに含まれる言語情報は，学習に用いた音声データから得られた言語の統計量であるため，“暗黙の言語情報” ということになる．Shallow Fusion では，“暗黙の言語情報” について考慮していない．テストドメインとトレーニングドメインが同じである場合には，Seq2Seq end-to-end 音声認識は，“暗黙の言語情報” を利用することで認識精度が向上する．しかし，テストドメインとトレーニングドメインが全く異なる場合は認識精度が低下する可能性がある．

提案手法は，ベイズ則に基づき，ASR モデルに含まれる “暗黙の言語情報” を確率的に除去する．ASR モデルに含まれている “暗黙の言語情報” は，ソースドメインのテキストデータを用いて学習された言語モデルで近似できると仮定する．“暗黙の言語情報”’’は, 式 (8) のように，ソースドメインの ASR モデルの出力確率から，ソースドメインのテキストデータを用いて学習した言語モデルの出力確率を差し引くことで除去することができる．

$$\begin{aligned}&\log P_{source}(y|x) - \lambda_{sub} \log \tilde{P}_{source}(y) \\ &\quad \propto \log P_{source}(x|y) + \log P_{source}(y) - \lambda_{sub} \log \tilde{P}_{source}(y) \\ &\quad \approx \log P_{source}(x|y)\end{aligned} \tag{8}$$

ここで，$\log \tilde{P}_{source}(y)$ はソースドメインのテキストデータで学習された言語モデルの出力確率であり，λ_{sub} は言語モデルの減算重みである．減算重みによって P_{source} と $\tilde{P}_{source}$ の推定誤差を補正することができる．式 (8) は，音響モデルの対数出力確率とみなすことができる．式 (8) にターゲットドメインのテキストデータで学習した言語モデルの出力確率を加えることで，言語情報を以下の式 (9) のように置き換えることができる．

$$\begin{aligned}\log P_{source}(y|x) - \lambda_{sub} \log \tilde{P}_{source}(y) + \lambda_{add} \log \tilde{P}_{target}(y) \\ \approx \log P_{source}(x|y) + \lambda_{add} \log \tilde{P}_{target}(y) \\ \propto \log P_{(source,target)}(y|x)\end{aligned} \tag{9}$$

ここで, $\log \tilde{P}_{target}(y)$ はターゲットドメインのテキストデータで学習された言語モデルの出力確率, λ_{add} は言語モデルの加算用重みである．$P_{(source,target)}(y|x)$ は音響情報がソースドメインに適応し，言語情報はターゲットドメインに適応したものであることを示す．つまり，$P_{(source,target)}(y|x)$ は音響情報がソースドメイン，言語情報がターゲットドメインの ASR モデルであると言える．

式 (6) から式 (9) より，ASR モデルの言語情報のみを特定ドメインの言語情報に置き換えることが可能であることを証明した．このように，任意のドメインのテキストデータを準備することで，そのタスクに対応した ASR モデルを作成することが可能である．また，提案法は Shallow Fusion のように推論時にのみ言語モデルを統合するため，2.1 節で述べた Cold Fusion や Deep Fusion のようなアプローチとは異なり，再学習は不要である．

表 1 実験に使用したデータセット情報

コーパス	データ分割	話者数	発話数	単語数
学術講演	Train	929	144,268	2,962,262
	Dev1	39	4,000	86,878
	Dev2	10	1,272	24,790
	Test	10	1,292	25,418
模擬講演	Train	1,654	232,886	3,340,546
	Dev1	38	4,018	71,239
	Dev2	10	1,385	16,325
	Test	13	1,336	24,157

表 2 学術講演および模擬講演コーパスを用いて学習した文字ベースの RNN 言語モデルの Test セットにおける PPL

Model	Test セットにおける PPL	
	学術講演	模擬講演
学術講演 LM	17.19	34.62
模擬講演 LM	33.29	18.87

4. 実験

4.1 実験条件

提案法である Language Model Replacement を ASR タスクで実験的に検証し，Shallow Fusion を用いたときの結果と比較した．テストセットの文字誤り率 (CER) を用いて評価した．ASR モデルには，2.2 節で紹介した Hybrid CTC/Attention Architecture を用いた．今回の実験では，2 種類の ASR モデルを準備した．1 つ目の ASR モデルは，工学，人文学および社会学の講演発表の音声データを含む学術講演コーパスを用いて学習した．2 つ目は，学術的な用語を含まない模擬講演コーパスを用いて学習した．いずれのコーパスも Corpus of Spontaneous Japanese (CSJ) データセットに含まれている．各コーパスに含まれる発話を，ランダムに Train セット，Dev1 セット，Dev2 セットおよび Test セットの 4 つに分割した．Train セットは ASR モデルの学習データ，Dev1 セットは ASR モデルの学習用の開発セット，Dev2 セットは Language Model Replacement および Shallow Fusion の言語モデル重みパラメータを調

表 3 言語モデル統合手法である Shallow Fusion (SF) と Language Model Replacement (LMR) の音声認識結果

統合手法	ASR モデル	Test セット	言語モデル統合情報	λ_{sub}	λ_{add}	CER
Baseline	学術講演	模擬講演	—	—	—	16.2
SF			$ASR_{\text{学術講演}} + \lambda_{add} LM_{\text{模擬講演}}$	—	0.4	15.3
LMR			$ASR_{\text{学術講演}} - \lambda_{sub} LM_{\text{学術講演}} + \lambda_{add} LM_{\text{模擬講演}}$	0.8	0.9	**14.0**
Baseline	模擬講演	学術講演	—	—	—	17.0
SF			$ASR_{\text{模擬講演}} + \lambda_{add} LM_{\text{学術講演}}$	—	0.3	15.4
LMR			$ASR_{\text{模擬講演}} - \lambda_{sub} LM_{\text{模擬講演}} + \lambda_{add} LM_{\text{学術講演}}$	0.8	0.8	**13.1**
Matched	学術講演	学術講演	—	—	—	9.8
	模擬講演	模擬講演	—	—	—	8.0

整するための開発セットとして用いた．また，テスト用のデータセットとして，Test set を使用した．すべてのデータセットは相互にオープンになっている．表 (1) に実験に用いたデータセットの詳細を示す．

言語モデルには，2.2 節で概説した文字ベースの RNN 言語モデルを用いた．それぞれのコーパスの Train セットで学習した言語モデルを Test セットを用いて算出した PPL (Perplexity) を表 2 に示す．表 (2) より，2 つのコーパスに含まれる言語情報が異なることが確認できる．表 2 は学術講演 ASR モデルの語彙リストを用いて PPL を算出した結果であるが，模擬講演 ASR モデルの語彙リストを用いた場合も同様の結果が得られた．

4.2 学習

ASR モデルへの入力音声の音響特徴量として 80 メルスケールのフィルタバンクを用い，また CMV 正規化を行った．さらに Train セットのデータ拡張として，話速変換を行い，元の音声データに 0.9 倍および 1.1 倍で変換した音声データを加えた．ASR モデルには，2.2 節で概説した Hybrid CTC/Attention Architecture を使用した．バッチサイズを 24 に設定し，最適化関数に Adadelta を用いて End-to-end で学習した．過学習防止のため，Dev1 セットを用いて Early Stopping を行った．この時，patience 値は 3 とした．2.2 節で説明したように，文字ベースの RNN 言語モデルは Embedding 層，2 層の LSTM 層，そして 1 層の Linear 層で構成されている．Embedding 層への入力は，文字の One-hot ベクトルとした．バッチサイズは 256 とし，最適化関数には SGD を使用した．

4.3 実験結果

学術講演および模擬講演 ASR モデルを用いて，Shallow Fusion (SF) と Language Model Replacement (LMR) の有用性を評価した．これらの言語モデル統合手法を学術講演および模擬講演 Test セットを用いて評価した．Dev2 セットを用いて，SF の言語モデルの加算重み λ_{add}，および提案法である LMR の言語モデルの加算重み λ_{add} 及び減算重み λ_{sub} を調整した．

実験結果を表 3 示す．また参考として，学術講演 ASR モデルを学術講演コーパスの Test セットで評価した際と，模擬講演 ASR モデルを模擬講演 Test セットで評価した際 (Matched case) の CER も示す．

学術講演 ASR モデルを模擬講演 Test セットを用いて評価した時，提案法である LMR を適用した時の CER は 14.0%であり SF における CER である 15.3%を上回る結果となった．また，模擬講演 ASR モデルを学術講演 Test セットを用いて評価した時，LMR を適用した時の CER は Shallow Fusion 適用時と比較して 2.3 ポイント改善した．このように，ASR モデルの学習データとテスデータのドメインが異なる場合，提案手法の性能は Shallow Fusion の性能を上回り，提案手法の有用性が示された．

また言語モデルの加算重み及び減算の重みは，提案手法を最適化するための重要なパラメータである．模擬講演ドメインにおいて，提案手法を学術講演 ASR モデルに適用した場合の減算重みと加算重みはそれぞれ 0.8 と 0.9 であった．LMR で使用した加算重みは Shallow Fusion で使用した加算重みよりも大きかったため，提案手法は対象ドメインの言語モデルをより効果的に活用できたことが示唆される．また学術講演言語モデルの出力確率の 0.8 倍という減算重みは，学術講演 ASR モデルに含まれる “暗黙の言語情報” の量を近似的に表していると考えられる．

5. 結論

End-to-end ASR モデルの学習は，音響情報と言語情報を同時に学習するため，言語情報のみを置き換えることは困難とされている．言語モデルと ASR モデルを統合する標準的な手法である Shallow Fusion では，ASR モデルに含まれる “暗黙の言語情報” が考慮されていない．

本稿では，ベイズ則に基づいて ASR モデルの “暗黙の言語情報” を除去し，ターゲットドメインの言語情報に置き換えることができる Language Model Replacement を提案した．ASR モデルの学習データで学習した言語モデルを用いて，ASR モデルに埋め込まれた “暗黙の言語情報” を減算することで除去した．そこにターゲットドメインの言語モデルの出力確率を加算し，言語情報を置き換えた．

実験では CSJ データセットの学術講演コーパスおよび模擬講演コーパスを用いて，ASR モデルに含まれる言語情報の置き換えが可能であることを実証した．学術講演コーパスで学習した ASR モデルを，模擬講演ドメインの Test セットで評価した場合，提案法である LMR の認識精度は Shallow Fusion での精度を上回り，CER を 1.3 ポイント改善した．また，模擬講演コーパスで学習した ASR モデルを，学術講演ドメインの Test セット使用して評価した場合，LMR は Shallow Fusion と比較し, CER を 2.3 ポイントを改善した．

謝辞 本研究は JSPS 科研費 JP19H01125, 18K18170 および 21K13641 の助成を受けたものです。

参考文献

[1] Lawrence Rabiner and Bing-Hwang Juan, Fundamentals of Speech Recognition, Prentice Hall, 1993.
[2] Dong Yu and Li Deng, Automatic Speech Recognition, Springer, 2015.
[3] Dzmitry Bahdanau, KyungHyun Cho and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate.” In ICLR, 2015.
[4] Jan K. Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho and Yoshua Bengio. “Attention-based models for speech recognition.” In Advances in Neural Information Processing Systems, pp. 577–585, 2015.
[5] Philipp Koehn. Statistical Machine Translation. Cambridge University Press, New York, NY, USA
[6] Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk and Yoshua Bengio. “On using monolingual corpora in neural machine translation.” arXiv preprint arXiv:1503.03535, 2015.
[7] Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh and Adam Coates. “Cold Fusion: Training Seq2Seq Models Together with Language Models.” arXiv preprint arXiv:1708.06426v1, 2017.
[8] Zhilin Yang, Bhuwan Dhingra, Ye Yuan, Junjie Hu, William W. Cohen and Ruslan Salakhutdinov. “Words or characters? Fine-grained gating for reading comprehension.” arXiv preprint arXiv:1611. 01724, 2016.
[9] Shinji Watanabe, Florian Boyer, Xuankai Chang, Pengcheng Guo, Tomoki Hayashi, Yosuke Higuchi, Takaaki Hori, Wen-Chin Huang, Hirofumi Inaguma, Naoyuki Kamo, Shigeki Karita, Chenda Li, Jing Shi, Aswin Shanmugam Subramanian and Wangyou Zhang. “The 2020 ESPnet update: new features, broadened applications, performance improvements, and future plans.” arXiv preprint arXiv:2012. 13006v1, 2020.
[10] Alex Graves, Santiago Fernandez, Faustino Gomez and Jurgen Schmidhuber. “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks.” In Proceedings of the 23rd International Conference on Machine Learning, pp. 369-376, 2006.
[11] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey and Tomoki Hayashi. “Hybrid CTC/Attention Architecture for End- to-End Speech Recognition.” In IEEE Journal of Selected Topics in Signal Processing, 2017.
[12] Tomas Mikolov. Statistical Language Models Based on Neural Networks. PhD thesis, Brno University of Technology, 2012.
[13] Frederick Jelinek, “Continuous speech recognition by statistical methods,” Proceedings of the IEEE, vol. 64, no. 4, pp. 532–556, 1976.
[14] Lalit R Bahl, Frederick Jelinek, and Robert L Mercer, “A maximum likelihood approach to continuous speech recognition,” IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 2, pp. 179–190, 1983.
[15] E. McDermott, H. Sak, and E. Variani, “A Density Ratio Approach to Language Model Fusion in End-to-End Automatic Speech Recognition,” ASRU2019, pp. 434-441, 2019.
[16] Z. Gong, N. Minematsu, D. Saito, “Language Model Augmentation in End-to-End ASR Systems Based on Noisy Channel Model,” Acoustical Society of Japan Spring Annual Meetings, (in Japanese), 2021.
[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J.,Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin,I. “Attention is All you Need,Advances in Neural Infor-mation Processing Systems(Guyon, I., Luxburg, U. V.,Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S.and Garnett, R., eds.),” Curran Associates, Inc. (2017).
[18] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. “Conformer: Convolution-augmented Transformer for Speech Recognition,” Interspeech 2020.