Vol.2020-NL-246 No.2
Vol.2020-SLP-134 No.2
2020/12/2

BERT による Sequence-to-Sequence 音声認識への知識蒸留

二見 颯[1,a)]　稲熊 寛文[1]　上乃 聖[1]　三村 正人[1]　坂井 信輔[1]　河原 達也[1]

概要：近年，音声認識において Sequence-to-Sequence (Seq2Seq) モデルが注目されている．Seq2Seq 音声認識モデルは音声とテキストのペアデータから学習されるため，テキストデータを追加利用することが課題となっている．また，Seq2Seq モデルではある単語を予測するためにそれ以前の文脈が用いられ，以後の文脈を用いることができない．そこで，本研究では BERT をある単語の以前だけでなく以後の文脈を用いて予測を行う言語モデルとして Seq2Seq 音声認識へ適用する．適用法として BERT を教師モデル，Seq2Seq 音声認識モデルを生徒モデルとした知識蒸留法を提案する．ここで，BERT の入力として複数発話にまたがるコンテキストを利用する． 主に日本語話し言葉コーパス (CSJ) 上の評価実験によって， 提案法による認識精度の大きな改善を確認した．さらに，従来の言語モデル適用法であるリスコアリングや Shallow Fusion と比較し，提案法は推論速度，認識精度ともに上回ることを確認した．

1. はじめに

近年，音響特徴列を単語列への写像を直接学習する End-to-End 音声認識が注目されている．End-to-End 音声認識の実現方法として，Connectionist Temporal Classification (CTC) [1]，RNN-Transducer [2]， Sequence-to-Sequence (Seq2Seq) モデル [3], [4], [5] を挙げることができる．本研究では，特に Seq2Seq モデルについて扱う．

Seq2Seq 音声認識では，学習に音声とテキストのペアデータが必要である．ただし，ペアデータの大規模な構築は容易ではないため，より入手が容易なテキストデータを追加利用できることが望ましい．そこで，テキストで学習された外部言語モデルを Seq2Seq モデルへ適用する方法がいくつか検討されている．リスコアリングでは，音声認識結果として得られた N-best 仮説を言語モデルで再評価し，最もスコアの高い仮説を認識結果として選択する．Shallow Fusion [6] や Cold Fusion [7] では，音声認識モデルのデコーダと外部言語モデルを組み合わせてデコードする．Shallow Fusion では推論時，Cold Fusion では学習時および推論時に言語モデルを用いる．最近では，知識蒸留法 (Knowledge Distillation) [8] に基づく言語モデルの適用法が提案されている [9]．教師モデルである言語モデルの予測をソフトラベルとして，Seq2Seq 音声認識モデルの学習に用いる．以上に挙げた外部言語モデル適用法では，従来 n-gram や RNN 言語モデル，Transformer [10] 言語モデルが利用される．これらはある単語を予測するためにそれ以前の単語列に基づく単方向の推論を行うため，本研究では「単方向」言語モデルとよぶこととする．

これらに対して，本研究では BERT [11] を言語モデルとして Seq2Seq 音声認識へ適用することを考える．BERT はマスクした単語を前後の単語列から予測する Masked Language Model (MLM) を事前学習として行うため，双方向の推論が実現できる．Seq2Seq モデルは単方向の推論を行い，単語予測において以後に現れる単語列の情報を利用できないが，BERT を言語モデルとして用いることで以後の文脈の利用が期待できる．BERT によるリスコアリングは既に提案されている [12], [13] が，認識結果は単方向の推論によって得られた N-best 仮説に限られ，また推論速度の面で問題がある．Shallow Fusion や Cold Fusion では，推論時，デコード途中には以後の文脈はまだデコードされていないため利用できず BERT の適用は困難である．そこで，本研究では知識蒸留法による BERT の適用を考える．教師モデルである BERT は，以前の文脈だけでなく以後の文脈に基づいたソフトラベルを生成することができ，Seq2Seq モデルの学習に用いることができる．この方法では，BERT は学習時のみ用いられ，推論時は不要である．さらに本研究では，BERT によるソフトラベル生成の際に，発話内の文脈だけでなく発話を超えて続く文脈を利用することを提案する．

[1] 京都大学 大学院情報学研究科
Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606–8501, Japan
[a)] futami@sap.ist.i.kyoto-u.ac.jp

2. 関連研究

2.1 Sequence-to-Sequence 音声認識

Seq2Seq モデルは，エンコーダとデコーダの 2 つのネットワークから構成される．エンコーダは音響特徴列 $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_T)$ を特徴列 $\boldsymbol{H} = (\boldsymbol{h}_1, ..., \boldsymbol{h}_T)$ へ変換する．デコーダは各ステップごとにエンコーダの特徴列と以前に出力した単語列に基づいて次の単語を予測し，単語列 $\boldsymbol{y} = (y_1, ..., y_N)$ を得る．エンコーダとデコーダは RNN [3], [4] あるいは Transformer [5] により実装される．

Seq2Seq モデルにおいて，i 番目にある語 v を予測する確率は，

$$P_{ASR}^{(i,v)} = p(v \mid \boldsymbol{X}, \boldsymbol{y}_{<i}) \tag{1}$$

と表せる．$\boldsymbol{y}_{<i}$ は y_i の以前の文脈 $(y_1, ..., y_{i-1})$ を表す．

学習時には，以下に表す損失関数を最小化する．

$$\mathcal{L}_{ASR} = -\sum_{i=1}^{N}\sum_{v=1}^{V} \delta(v, y_i) \log P_{ASR}^{(i,v)} \tag{2}$$

$\delta(v, y_i)$ は $v = y_i$ のとき 1, $v \neq y_i$ のとき 0 の値をとる．

2.2 BERT

BERT [11] は，複数層の Transformer モデルによって構成される．BERT は大規模なラベルなしテキストを用いた事前学習法として提案され，少数のラベルありテキストによる fine-tuning によって，自然言語処理 (NLP) のさまざまなタスクで効果を示した．BERT 以前の事前学習法として OpenAI GPT [14] や ELMo [15] が挙げられるが，OpenAI GPT は単方向，ELMo は双方向だが left-to-right と right-to-left の RNN 言語モデルが最終層のみで「浅く」結合された構造である．一方 BERT では，Transformer により双方向の推論が 1 層目から「深く」結合されている．

BERT は事前学習タスクとして Masked Language Model (MLM) と Next Sentence Prediction (NSP) が行われる．MLM では，入力単語の一部をランダムに `[MASK]` トークンに置換し，前後の文脈から元の単語を予測する学習を行う．本研究では，BERT を前後の文脈を用いて単語予測を行う言語モデルとして用いることを目的とするため，MLM のみを行い，NSP および fine-tuning は行わない．

2.3 知識蒸留法

知識蒸留法 (Knowledge Distillation) [8] は Teacher-Student 学習ともよばれ，生徒 (student) モデルの出力分布を教師 (teacher) モデルの出力分布へ近似するよう学習させる．この方法は，主に大規模なパラメータを持つ教師モデルを小規模なパラメータを持つ同種の生徒モデルで近似するモデルの圧縮を目的として用いられる [16], [17].

一方，モデルの圧縮だけでなく教師モデルで学習された知識を別種の生徒モデルに転移することも行われる．特に [9] では，知識蒸留法によってテキストで学習された単方向言語モデルを Seq2Seq 音声認識へ適用した．また，これに続く研究として，Causal clOze completeR (COR) を Seq2Seq 音声認識へ適用することが提案された [18]．COR は left-to-right と right-to-left の Transformer モデルを最終層のみで結合した構造 [19] で，単語予測に前後の文脈を利用できる．しかし，COR は ELMo [15] 同様「浅い」双方向の結合である．本研究では「深い」結合であり，簡潔な構造である BERT を用いる．さらに本研究では，より良い教師分布を与えるため複数発話にまたがるコンテキストを利用する．機械翻訳の分野では，BERT にソース言語の文とターゲット言語の文を与えて得た教師分布による Seq2Seq モデルへの知識蒸留が提案されている [20].

3. 提案手法

本研究では，BERT を教師モデル，SeqSeq 音声認識モデルを生徒モデルとした知識蒸留法を提案する．BERT の生成したソフトラベルにより，Seq2Seq モデルはより文法的あるいは意味的に尤もらしい認識結果を出力するよう学習される．

$\boldsymbol{X}$ をある発話の音響特徴列，$\boldsymbol{y}$ を $\boldsymbol{X}$ に対応する現発話の単語列とする．本研究では，BERT において複数発話にまたがるコンテキスト，つまり前後の発話を利用する．$\boldsymbol{y}^{(L)} = (y_1^{(L)}, ..., y_L^{(L)})$ を現発話以前の発話に含まれる単語列，$\boldsymbol{y}^{(R)} = (y_1^{(R)}, ..., y_R^{(R)})$ を現発話以後の発話に含まれる単語列とする．それぞれの長さ L，R は，現発話の長さ N と合わせて一定 (例: $L + R + N = 256$) かつ $L = R$ となるよう決定する．

発話内に限らず，複数発話にまたがるコンテキストを利用する利点として 2 つ挙げることができる．第一に，教師分布をより正確にできることが挙げられる．特に短い発話では，発話内の文脈から単語を予測することが困難になる．発話外のコンテキストを追加することで，発話の長さによらずより正確な教師分布を Seq2Seq モデルの学習に用いることができる．第二に，BERT の学習時と推論時の入力系列長の不一致を解決できることが挙げられる．BERT の学習時は，一定の系列長 (例: 256) の入力を用いる [21] が，発話の長さはさまざまである．前後の発話に含まれる単語を追加することで，学習時と同じ系列長の入力によって BERT の推論を行うことができる．

BERT において，i 番目にある語 v を予測する確率は以下に表される．

$$P_{BERT}^{(i,v)} = p(v \mid [\boldsymbol{y}^{(L)}; \boldsymbol{y}_{\setminus i}; \boldsymbol{y}^{(R)}]) \quad = \frac{\exp(z_v/T)}{\sum_{j=1}^{V} \exp(z_j/T)} \tag{3}$$

z_j は BERT の出力，T は分布を調整するハイパーパラメー

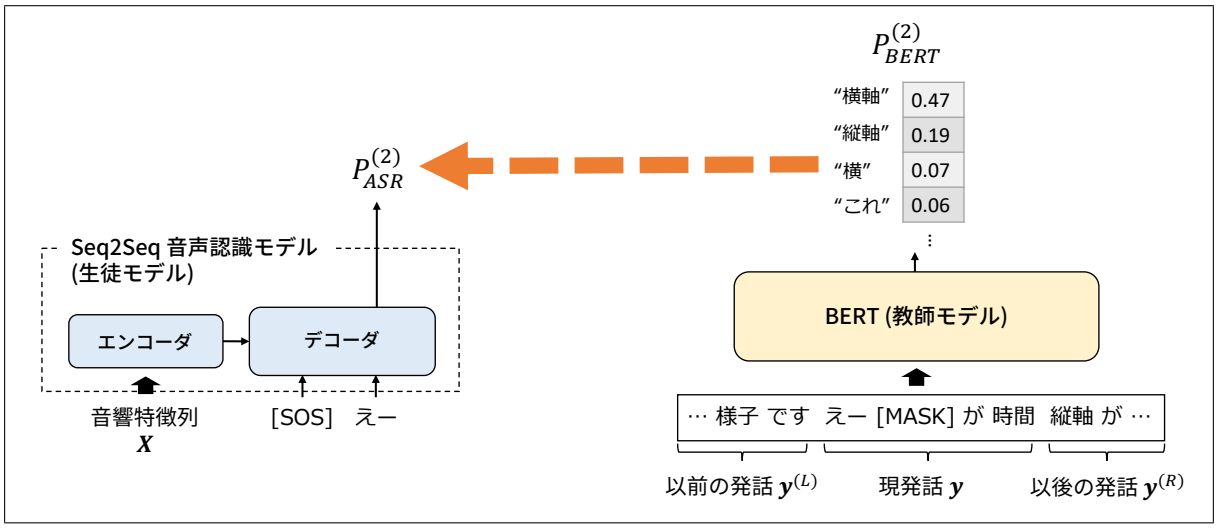

図 1 提案法の概要．$P^{(2)}_{BERT}$ は，[MASK] に置換した y_2 (この例では「横軸」) を現発話外を含めた前後の文脈から予測することで得る．Seq2Seq モデルの学習では正解ラベルだけでなく，このソフトラベルも用いられる．

タである．$\boldsymbol{y}_{\setminus i}$ は $\boldsymbol{y}$ の i 番目の単語を [MASK] に置換した $(y_1, ..., y_{i-1}, \texttt{[MASK]}, y_{i+1}, ..., y_N)$ を表す．$\boldsymbol{y}_{\setminus i}$ は $\boldsymbol{y}^{(L)}$, $\boldsymbol{y}^{(R)}$ と結合され，BERT へ入力される．

$P^{(i)}_{ASR}$, $P^{(i)}_{BERT}$ はそれぞれ Seq2Seq モデルと BERT による i 番目の予測分布を表す．知識蒸留法では，図 1 に示すように，各 i に対して $P^{(i)}_{ASR}$ と $P^{(i)}_{BERT}$ の KL ダイバージェンスを最小化する．

$$KL(P^{(i)}_{BERT} || P^{(i)}_{ASR}) = -\sum_{v=1}^{V} P^{(i,v)}_{BERT} \log \frac{P^{(i,v)}_{ASR}}{P^{(i,v)}_{BERT}} \quad (4)$$

$P^{(i,v)}_{BERT}$ は固定されることから，KL ダイバージェンスによる損失関数は，

$$\mathcal{L}_{KD} = -\sum_{i=1}^{N}\sum_{v=1}^{V} P^{(i,v)}_{BERT} \log P^{(i,v)}_{ASR} \quad (5)$$

と表される．

提案法では，Seq2Seq モデルの学習時に $\mathcal{L}_{ASR}$ と $\mathcal{L}_{KD}$ を組み合わせた損失関数を用いる．

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{ASR} + \alpha\mathcal{L}_{KD} \ \ (0 \le \alpha \le 1) \quad (6)$$

これは正解 (ハード) ラベルの one-hot 表現 $\delta(v, y_i)$ とソフトラベル $P^{(i,v)}_{BERT}$ を用いて，以下のように表される．

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{v=1}^{V}((1-\alpha)\delta(v, y_i) + \alpha P^{(i,v)}_{BERT}) \log P^{(i,v)}_{ASR} \quad (7)$$

$P^{(i)}_{BERT}$ は，学習データに関して事前計算できる．そこで，メモリ効率のため 上位 K 語の確率分布のみ利用する top-K distillation [22] を適用する．本研究では $K = 8$ と設定した．BERT は推論に時間を要するが，事前計算により Seq2Seq モデルの学習時に BERT の推論を行う必要がなく，また Seq2Seq モデルの推論時にも BERT を必要としない．

4. 評価実験

4.1 実験設定

本研究では日本語のデータセットとして，「日本語話し言葉コーパス」(CSJ) [23] および「日本語書き言葉均衡コーパス」(BCCWJ) [24] を用いた．CSJ は学会講演を収録した CSJ-APS (約 240 時間) と模擬講演を収録した CSJ-SPS (約 280 時間) のサブコーパスで構成される．CSJ-APS は音声認識の学習，BCCWJ (約 77M 単語) は，CSJ-APS (約 3.9M 単語)，CSJ-SPS (約 4.1M 単語) とともに言語モデルの学習に用いた．評価セットは CSJ-eval1 とした．音声認識と言語モデルで語彙を共有し，Byte Pair Encoding [25] による語彙サイズ 7520 のサブワード単位を用いた．また，英語のデータセットとして TED-LIUM2 コーパス [26] および Europarl (v7) コーパス [27] を用いた．TED-LIUM2 (約 210 時間) は音声認識の学習，Europarl (約 106M 単語) は TED-LIUM2 (約 3.5M 単語) とともに言語モデルの学習に用いた．語彙サイズは 1062 とした．

Seq2Seq 音声認識モデルは層数 L=5, 隠れ状態数 H=320 の双方向 LSTM によるエンコーダ，L=1，H=320 の単方向 LSTM によるデコーダから構成した．モデルは学習率 1e-3 の Adam [28] アルゴリズムにより最適化した．正則化手法として，ハードラベル，ソフトラベルともに $p = 0.1$ の label smoothing [29] を適用した．特にソフトラベルに関しては，top-K ($K = 8$) 以外の単語へ $p = 0.1$ を割り当てた．また，データ拡張手法として SpecAugment [30], TED-LIUM2 では speed perturbation [31] も用いた．推論時には，ビーム幅 5 のビームサーチを行った．

言語モデルは，BERT と単方向言語モデルを L=6,

表 1 CSJ-APS における知識蒸留法による言語モデル適用の効果．α は式 (7) に示されるソフトラベルの重み，LM-Acc は学習データに対するソフトラベルにおいて最尤の単語が正解単語と一致する確率を表す．

言語モデル	コンテキスト	WER(%)	α	LM-Acc(%)
なし	-	10.31	0	-
uniLM	発話内	9.88	0.05	45.1
uniLM	256	10.01	0.1	55.1
BERT	発話内	9.53	0.2	64.6
BERT	256	**9.19**	0.3	77.2

H=512，ヘッド数 A=8 の Transformer から構成し，同じデータで学習した．また，学習時の入力系列長は 256 とし，BERT の学習では入力の 8 % を [MASK] に置換した．モデルは Adam アルゴリズムにより最適化し，学習率ははじめ 10% は 1e-4 まで増加させ (warmup)，その後線形に減少させた．

4.2 結果

表 1 に CSJ-APS における，知識蒸留法による言語モデル適用の効果を示す．ここで言語モデルは単方向の Transformer 言語モデル (uniLM) と BERT，さらにそれぞれソフトラベル生成に発話外のコンテキストを利用するかどうかを比較した．表 1 中の「発話内」については発話内の文脈のみ，「256」については， uniLM では以前の発話，BERT は前後の発話に含まれる単語を 256 単語になるまで結合し uniLM，BERT の入力としたことを表す．表 1 の通り，知識蒸留法を用いた音声認識によって単語誤り率 (WER) の改善が見られたが，言語モデルとして uniLM より前後の文脈を利用できる BERT を用いた場合の改善が大きい．さらに，BERT では発話外のコンテキストを利用する効果が見られ，提案法では言語モデルなしのベースラインから相対的に 10.9% の WER の改善がみられた．表 1 中の LM-Acc は学習データに対する言語モデルの予測において，最尤の単語が正解単語と一致する確率，すなわちソフトラベルの正確さを表す．単語予測に以前の文脈に加えて以後の文脈，さらに発話を超えて続く文脈を利用することで，LM-Acc に示すようにソフトラベルをより正確にすることができた．式 (7) のソフトラベルの重み α は開発セットの WER が最小になるものを選択したが，正確なソフトラベルにより α を大きくすることができた．

表 2 に，発話外のコンテキストを利用する効果について詳細に分析する．CSJ-APS の発話の長さは平均約 24 単語，最小のもので 1 単語，最大のもので 118 単語であった．BERT の学習時，蒸留つまりソフトラベルの推論時にどれだけコンテキストを利用するか変化させ，CSJ-APS における音声認識の WER を比較した．まず，BERT は学習時により長いコンテキストを見て単語間の関係を学習することが重要であった．また，蒸留時は学習時と一致する系列長になるまでコンテキストを利用する場合が最適であった．

表 2 BERT の学習，蒸留 (推論) 時の単語予測におけるコンテキスト利用の比較．

コンテキスト		
BERT 学習時	蒸留時	WER(%)
64	発話内	9.91
64	64	9.69
128	発話内	9.62
128	128	9.40
256	発話内	9.53
256	64	9.28
256	128	9.28
256	256	**9.19**

表 3 CSJ-APS における従来の言語モデル適用法との比較．PPL は認識結果の uniLM, BERT によるパープレキシティを表す．

		PPL	
	WER(%)	uniLM	BERT
ベースライン (LM なし)	10.31	35.7	7.9
+ リスコアリング (uniLM)	9.66	30.6	6.5
+ Shallow Fusion (uniLM)	9.79	32.0	6.8
+ リスコアリング (BERT)	9.61	31.2	6.2
+ 提案法	**9.19**	32.4	6.7

表 3 に提案法を従来の言語モデル適用法であるリスコアリング，Shallow Fusion と比較する．リスコアリングは音声認識の 50-best 仮説に対して uniLM，BERT を用い，Shallow Fusion はビーム幅 5 で uniLM を用いた．表 3 の通り，提案法による WER の改善は従来の言語モデル適用法を上回るものであった．さらに，提案法は推論時に言語モデルが不要であり，他の適用法と比較し高速に推論できる．それぞれの認識結果について uniLM，BERT によるパープレキシティ (PPL) を比較した．特に，BERT による PPL は仮説 $\boldsymbol{w} = (w_1, ..., w_N)$ に対して以下に定義する [32] .

$$PPL(\boldsymbol{w}) := \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log p(w_i|\boldsymbol{w}_{\setminus i})\right) \qquad (8)$$

提案法により，ベースラインと比較して PPL の改善が見られたが，提案法では学習時に言語モデルの知識を蒸留するため，推論時に直接適用するリスコアリングおよび Shallow Fusion ほどの PPL の改善は見られなかった．

図 2 に，ベースラインと提案した知識蒸留法を適用した場合の学習経過を示す．提案法は特に学習初期 (図左) において効果を示し，その後も提案法はベースラインを下回る WER で推移することがわかった．

表 4 に英語コーパス TED-LIUM2 において提案法を適用した効果を示す．提案法により，TED-LIUM2 においてもベースラインから相対的に 11.8% の WER の改善が見られた．

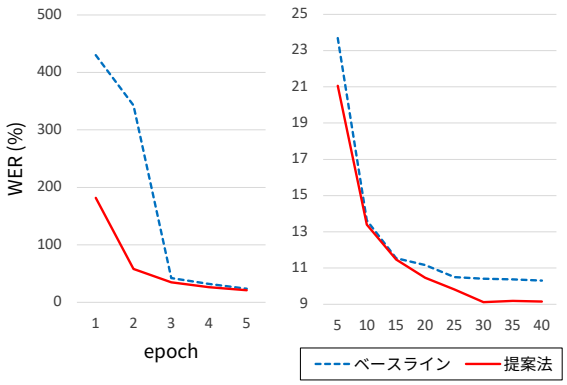

図 2 提案法による Seq2Seq 音声認識モデルの学習経過．各 epoch ごとの WER を表す．

表 4 TED-LIUM2 コーパスにおける提案法の効果．

	WER(%)	α	LM-Acc(%)
ベースライン	11.23	0	-
+提案法	**9.90**	0.3	82.0

5. おわりに

BERT はテキストデータから学習され，Seq2Seq モデルが利用できない前後の文脈を用いて単語を予測することができる．本研究では，BERT を教師モデル，Seq2Seq 音声認識モデルを生徒モデルとした知識蒸留法によって，BERT を言語モデルとして音声認識へ適用することを提案した．ここで，発話を超えて続く文脈に基づく教師分布を利用した．評価実験により，単方向言語モデルや従来の言語モデル適用法に対する提案法の有効性を示すことができた．今後の研究として，言語モデルとして XLNet [33] や ELECTRA [34] の音声認識への適用や， BERT の CTC [1] や RNN-Transducer [2] への適用を考えている．

参考文献

[1] Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J.: Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks, *ICML*, pp. 369–376 (2006).
[2] Graves, A.: Sequence Transduction with Recurrent Neural Networks, *arXiv* (2012).
[3] Chan, W., Jaitly, N., Le, Q. and Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, *IEEE ICASSP*, pp. 4960–4964 (2016).
[4] Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y.: Attention-Based Models for Speech Recognition, *NeurIPS*, pp. 577–585 (2015).
[5] Dong, L., Xu, S. and Xu, B.: Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition, *IEEE ICASSP*, pp. 5884–5888 (2018).
[6] Chorowski, J. and Jaitly, N.: Towards Better Decoding and Language Model Integration in Sequence to Sequence Models, *Interspeech*, pp. 523–527 (2017).
[7] Sriram, A., Jun, H., Satheesh, S. and Coates, A.: Cold Fusion: Training Seq2Seq Models Together with Language Models, *Interspeech*, pp. 387–391 (2018).
[8] Hinton, G. E., Vinyals, O. and Dean, J.: Distilling the Knowledge in a Neural Network, *arXiv* (2015).
[9] Bai, Y., Yi, J., Tao, J., Tian, Z. and Wen, Z.: Learn Spelling from Teachers: Transferring Knowledge from Language Models to Sequence-to-Sequence Speech Recognition, *Interspeech*, pp. 3795–3799 (2019).
[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, *NeurIPS*, pp. 5998–6008 (2017).
[11] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *NAACL*, pp. 4171–4186 (2019).
[12] Shin, J., Lee, Y. and Jung, K.: Effective Sentence Scoring Method Using BERT for Speech Recognition, *Proceedings of The Eleventh Asian Conference on Machine Learning*, pp. 1081–1093 (2019).
[13] Salazar, J., Liang, D., Nguyen, T. Q. and Kirchhoff, K.: Masked Language Model Scoring, *ACL*, pp. 2699–2712 (2020).
[14] Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I.: Improving language understanding by generative pre-training, *OpenAI technical report* (2018).
[15] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L.: Deep Contextualized Word Representations, *NAACL*, pp. 2227–2237 (2018).
[16] Kim, Y. and Rush, A. M.: Sequence-Level Knowledge Distillation, ACL, pp. 1317–1327 (2016).
[17] Sanh, V., Debut, L., Chaumond, J. and Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, *arXiv* (2019).
[18] Bai, Y., Yi, J., Tao, J., Tian, Z., Wen, Z. and Zhang, S.: Integrating Whole Context to Sequence-to-sequence Speech Recognition, *arXiv* (2019).
[19] Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L. and Auli, M.: Cloze-driven Pretraining of Self-attention Networks, *EMNLP-IJCNLP*, pp. 5360–5369 (2019).
[20] Chen, Y.-C., Gan, Z., Cheng, Y., Liu, J. and Liu, J.: Distilling Knowledge Learned in BERT for Text Generation, *ACL*, pp. 7893–7905 (2020).
[21] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach, *arXiv* (2019).
[22] Tan, X., Ren, Y., He, D., Qin, T. and Liu, T.-Y.: Multilingual Neural Machine Translation with Knowledge Distillation, *ICLR* (2019).
[23] Maekawa, K.: Corpus of Spontaneous Japanese : its design and evaluation, *SSPR* (2003).
[24] Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M. and Den, Y.: Balanced Corpus of Contemporary Written Japanese, *Lang. Resour. Eval.*, pp. 345–371 (2014).
[25] Sennrich, R., Haddow, B. and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725 (2016).
[26] Rousseau, A., Deléglise, P. and Estève, Y.: Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks, *LREC*, pp. 3935–3939

(2014).
[27] Koehn, P.: EuroParl: A parallel corpus for statistical machine translation, *MT Summit*, pp. 79–86 (2004).
[28] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *CoRR* (2015).
[29] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z.: Rethinking the Inception Architecture for Computer Vision, *CVPR*, pp. 2818–2826 (2016).
[30] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D. and Le, Q. V.: SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition, *Interspeech*, pp. 2613–2617 (2019).
[31] Ko, T., Peddinti, V., Povey, D. and Khudanpur, S.: Audio augmentation for speech recognition, *Interspeech*, pp. 3586–3589 (2015).
[32] Chen, X., Ragni, A., Liu, X. and Gales, M. J.: Investigating Bidirectional Recurrent Neural Network Language Models for Speech Recognition, *Interspeech*, pp. 269–273 (2017).
[33] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R. and Le, Q. V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding, *NeurIPS*, pp. 5753–5763 (2019).
[34] Clark, K., Luong, M.-T., Le, Q. V. and Manning, C. D.: ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, *ICLR* (2020).