4Q-02

双方向 Transformer 言語モデルによる音声認識仮説のリスコアリング

二見 颯[1]　上乃 聖[2]　三村 正人[2]　坂井 信輔[2]　河原 達也[2]

[1] 京都大学 工学部情報学科　[2] 京都大学 大学院情報学研究科

1. はじめに

音声認識は従来，音響信号から音韻情報を抽出する音響モデル，発音辞書，単語系列の尤度を計算する言語モデルに分け，それぞれ学習して統合する DNN-HMM 音声認識モデルで実現されてきた．近年ではそれらを一つのモデルで実現する End-to-End 音声認識も提案されている．両モデルでは一度音声認識仮説を出力した後に，テキストデータのみで学習された外部言語モデルを用いて，仮説のスコアを計算し直すリスコアリングが効果を示している [1].

本研究では，言語モデルとして双方向 Transformer [2] を用いることを考える．双方向 Transformer は，順方向の推論と逆方向の推論の出力を単に結合した双方向の RNN とは異なり，自己注意機構により双方向の長い文脈を直接捉えることができる． モデル構造として双方向 Transformer を採用した BERT [3] はマスクされた単語を前後の文脈から予測する事前学習 (Masked Language Model) を行うことで，自然言語処理の様々なタスクで効果を示してきた．さらに，音声認識仮説のリスコアリングにおいても双方向 Transformer 言語モデルが検討されている [4].

2. リスコアリング

$\boldsymbol{W} = w_1 w_2 ...$ を単語列，$\boldsymbol{X}$ を音響特徴量とする．リスコアリングでは，音声認識モデルの対数尤度 $Score_{ASR} := \log p(\boldsymbol{W}|\boldsymbol{X})/|\boldsymbol{W}|$ と言語モデルのスコア $Score_{LM}$ を組み合わせて

$$Score := (1-\lambda) Score_{ASR} + \lambda Score_{LM} \quad (1)$$

が最大となる仮説を N-best 仮説の中から選択する．

従来用いられる言語モデルとして，N グラム，単方向 RNN 言語モデル等があり，これら単方向言語モデルによるスコアは同時確率

$$Score_{LM} := \sum_{t=1}^{|\boldsymbol{W}|} \log p(w_t|\boldsymbol{W}_{<t}) = \log p(w_1, ..., w_{|\boldsymbol{W}|}) \quad (2)$$

によって計算される．

一方，双方向 Transformer 言語モデルのスコアは

$$Score_{LM} := \sum_{t=1}^{|\boldsymbol{W}|} \log p(w_t|\boldsymbol{W}_{\backslash t}) \quad (3)$$

によって計算される [4]. $\boldsymbol{W}_{\backslash t}$ は $\boldsymbol{W}$ の t 番目の単語を [MASK] トークンに置き換えたものである．双方向 Transformer 言語モデルの学習は，[MASK] トークンから元の単語を予測できるように，1 発話の 15 % をランダムに選び [MASK] トークンに置き換える Masked Language Model によって実現される [4].

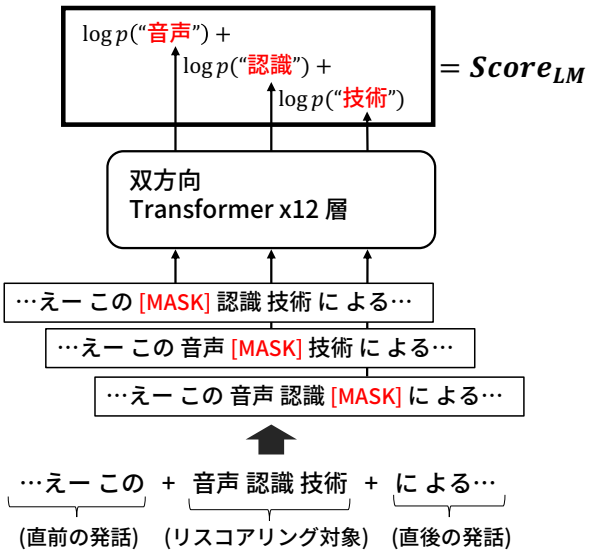

図 1: 双方向 Transformer 言語モデルによるスコア計算．リスコアリング対象「音声認識技術」の単語を 1 語ずつ [MASK] に置き換えた入力から元の単語を予測することで言語モデルのスコアを得る．

3. 提案手法

3.1 複数発話の結合

Transformer は自己注意機構により，長い文脈を捉えることができる [2]．前後の発話の音声認識仮説を結合しスコア計算を行うことで，現発話の音声認識仮説内で捉えることができない文脈を考慮できると考えられる．本研究では，学習・リスコアリング時ともに発話の前後 1 発話を結合し，計 3 発話分を入力とする．リスコアリング時には，結合する前後の発話は音声認識の対数尤度が最大のものを使用する．図 1 に「音声認識技術」という発話へのスコア計算を例示する．

3.2 スムージング

マスクされた単語を前後の文脈から予測する双方向モデルでは， 単方向モデルに比べ，より学習データのテキストの影響を受けやすく，パープレキシティが下がりやすいと考えられる．しかし，パープレキティが低すぎるモデルでは，フィラーや音声認識誤りが含まれるリスコアリング時と学習時のデータの不一致により，正しくリスコアリングができないと考えられる．そのため，本研

Rescoring Hypotheses of Automatic Speech Recognition with Bidirectional Transformer Language Model: *Hayato Futami, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara (Kyoto Univ.)*

究では [5] を基に，以下の式 (4) に示すソフトマックス層でのスムージングを適用する．

$$p(w_t|\boldsymbol{W}_{\setminus t}) = \frac{\exp(\alpha y_{w_t})}{\sum_j \exp(\alpha y_j)} \quad (4)$$

y_j をソフトマックス層への入力，$\alpha(<1)$ をスムージングのパラメータとする．

4. 評価実験

本研究ではデータセットとして『日本語話し言葉コーパス』(CSJ) を用いる．CSJ は CSJ-APS と CSJ-SPS の 2 つのサブコーパスで構成されている．このうち，CSJ-APS の音声とテキストのペアを音声認識モデルの学習，CSJ-APS と CSJ-SPS のテキストを言語モデルの学習に用いた．提供されている CSJ-APS 用のテストセットのうち，テストセット 2 を開発データ，テストセット 1 を評価データとした．(3) の言語モデルのスコアの重み λ は (0.0025, 0.005, ..., 0.05) から評価データの WER を最小にするものを示す．式 (4) のスムージングのパラメータ α は開発データにより $\alpha = 0.5$ と決定する．語彙は音声認識モデル，言語モデルともに Byte Pair Encoding [6] を用いて語彙サイズ 12414 のサブワード単位とする．

音声認識モデル　音響特徴量からサブワード系列を End-to-End で推定するモデルとして，注意機構モデル [7] を用いる．注意機構モデルは 5 層の双方向 LSTM エンコーダと 1 層の単方向 LSTM デコーダで構成される．正則化手法としてラベルスムージング，データ拡張手法として SpecAugment [8] を用いる．

双方向 Transformer 言語モデル　双方向 Transformer 言語モデルは，$\mathrm{BERT}_{\mathrm{BASE}}$ [3] と同様隠れ状態 768 次元，12 個の注意機構ヘッドを持つ 12 層の双方向 Transformer とする．

比較手法の N グラムは 4 グラム (4-gram)，RNN 言語モデル (RNNLM) は，512 次元の出力を持つ埋め込み層および隠れ状態 512 次元を持つ単方向 LSTM 3 層により構成される．評価データの音声認識 50-best 候補に対するリスコアリング後の単語誤り率 (WER) によって評価を行った．

表 1 に先行研究 [4] (biTLM) の双方向 Transformer 言語モデルと提案した前後 1 発話を追加したもの (biTLM(+前後発話)) とさらにスムージングを適用したもの (biTLM(+前後発話、スムージング)) との比較結果を示す．ともに効果が確認できる．

表 2 に提案した双方向 Transformer 言語モデルと N グラム，RNN 言語モデルとの比較結果を示す．前後発話を追加した双方向 Transformer 言語モデルとの公正な比較のため，N グラム言語モデルと RNN 言語モデルにおいても前 2 発話を追加してリスコアリング結果の WER を評価した．

表 2 では評価データ上のパープレキシティ (PPL) についても比較している．双方向言語モデルのパープレキシティについては，式 (5) で定義される擬似パープレキシティ [5] とする．

$$PPL(\boldsymbol{W}) := \frac{1}{|\boldsymbol{W}|}\sum_{t=1}^{|\boldsymbol{W}|} -\log p(w_t|\boldsymbol{W}_{\setminus t}) \quad (5)$$

スムージングの根拠とした双方向言語モデルの予測単語のスパースな確率分布は低いパープレキシティとして現れている．また単方向言語モデルに対しても同様のスムージングを行ったが効果は見られなかった．

表 1: 提案手法によるリスコアリングの効果

	WER(%)
音声認識のみ	10.57
biTLM [4]	9.96
biTLM (+前後発話)	**9.67**
biTLM (+前後発話，スムージング)	**9.52**
50-best oracle	5.41

oracle とは仮説の中で全て WER を最小にするものを選べた場合の WER であり，リスコアリングによる WER 改善の下限である．

表 2: 単方向言語モデルとの比較 (3 発話)

	WER(%)	PPL
音声認識のみ	10.57	—
4-gram	10.14	83.43
RNNLM	9.92	69.36
biTLM (+前後発話)	**9.67**	9.51
biTLM (+前後発話，スムージング)	**9.52**	19.06

5. おわりに

本稿では，評価実験によって，双方向 Transformer 言語モデルが音声認識仮説のリスコアリングに有効であることを示し，さらに提案した複数発話の結合とスムージングの有効性を示した．今後は，より大規模なテキストデータでの学習，さらには，デコード段階で双方向の文脈を考慮できるようにすることで N-best 仮説として現れない正解をデコードすることを目指す．

参考文献

[1] W. Chan *et al.*: "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition" *ICASSP*, 4960–4964, 2016.

[2] A. Vaswani *et al.*: "Attention Is All You Need" *NIPS*, 5998–6008, 2017.

[3] J. Devlin *et al.*: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" *NAACL*, 4171–4186, 2019.

[4] J. Shin *et al.*: "Effective Sentence Scoring Method Using BERT for Speech Recognition" *ACML*, 1081–1093, 2019.

[5] X. Chen *et al.*: "Investigating Bidirectional Recurrent Neural Network Language Models for Speech Recognition" *INTERSPEECH*, 269–273, 2017.

[6] R. Sennrich *et al.*: "Neural Machine Translation of Rare Words with Subword Units" *ACL*, 1715–1725, 2016.

[7] J. Chorowski *et al.*: "Attention-Based Models for Speech Recognition" *NIPS*, 577–585, 2015.

[8] DS. Park *et al.*: "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition" *INTERSPEECH*, 2613–2617, 2019.