6Q-07

相槌・フィラー予測とのマルチタスク学習による円滑なターンテイキング

原 康平 † 井上 昂治 ‡ 高梨 克也 ‡ 河原 達也 ‡

† 京都大学 工学部情報学科 ‡ 京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

人間との自然なインタラクションを指向した対話システムの研究・開発が盛んになっている．自然な対話を実現するためには，円滑なターンテイキングの実現が必要である．二者間の対話におけるターンテイキングの予測とは，ある発話単位の末尾において，どちらの対話参与者が次の発話権を得るか (話者交替または発話権保持) を予測する問題である．ターンテイキングの予測モデルとして，Support Vector Machine (SVM) [1] や Long Short-Term Memory (LSTM) [2, 3] が用いられている．また，特徴量としては先行発話の韻律情報や言語情報などが用いられている．

本稿では，ターンテイキングの予測と他のふるまいとの関係性を考慮する．ターンテイキングに関係するふるまいとしては相槌とフィラーが挙げられる．相槌とは，「あー」や「うん」などの聞き手応答を指し，現話者の発話権保持を促すなどの役割がある [4]．フィラーとは，「えー」や「あのー」といった言い淀み時などに出現する場繋ぎ的な表現を指し，適切な言語表現の選択や発話権の取得・保持などのためにも用いられる [5]．つまり，相槌には相手発話の継続を促す役割があり，フィラーには発話権を獲得する意志があることを相手に示唆する役割がある．したがって，これらのターンテイキングに関わるふるまいとの関係性を考慮することによって，ターンテイキングの予測精度も向上することが期待される (図 1)．本研究では，ターンテイキングと相槌・フィラーの予測について，マルチタスク学習による統合モデルを提案する．

2. ベースラインモデル

本研究におけるベースラインモデルでは，Skantze [2] のモデルを参考とする．このモデルは，中間層 1 層の LSTM で構成され，対話参与者毎にモデルを用意して，将来におけるその発話を予測する．各時間フレームで入力を受け取り，その時点から将来 20 フレーム（1 sec）分の区間においてその参与者が発話する確率を出力とする．つまり，出力は 20 次元で，各次元が将来の各時間フレームにおける発話確率に対応する．入力特徴量は，現フレームで話しているか否かの 2 値，基本周波数（F0）の絶対値および平均からの相対値，パワー，Spectral Stability である．なお，各参与者が話しているか否かの 2 値を除く各特徴量は参与者毎に平均 0，分散 1 に正規化されている．本研究では，上記に加え，F0 とパワーの Δ，ΔΔ も使用する．以上を参与者毎に抽出して 1 つに結合する．したがって，一対一での対話の場合，入力次元数は 18（=9 次元 ×2 参与者）である．フレームのシフトサイズは 50 msec である．本研究で使用するモデルでは，LSTM 層 (18 ユニット) に加え，全結合層 (20 ユニット) を 2 層分追加した．

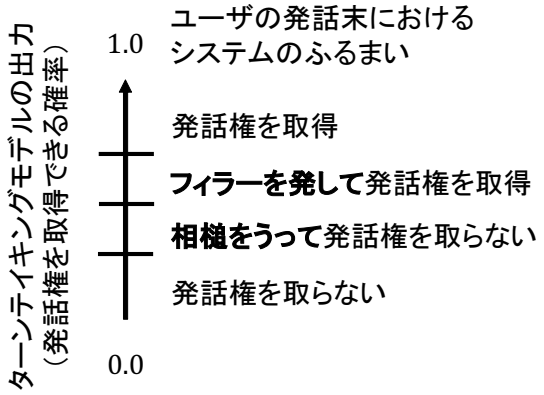

図 1: ターンテイキングと相槌・フィラーの関係

実際のターンテイキングの予測は，上述のモデルを用いてフレーム毎ではなく間休止単位（Inter-Pausal Unit, IPU）末でのみ行う．IPU とは，200 msec 以上のポーズで区切られた発話である．IPU 末であることを認定するために，発話末から 200 msec のポーズが経過した時点を予測時点とする．ただし，将来の 20 フレーム以内に両参与者ともに発話しない，または両参与者ともに発話する場合は予測対象外とする．ターンテイキングを予測するために，まず，予測対象時点での出力 (将来 1 sec 分の発話の確率) を両参与者のモデルから取得する．次に，出力である 20 次元の出力の平均を参与者ごとに算出し，値が大きい方の参与者を次話者とする．ここで，先行話者と次話者が異なる場合は話者交替 (take)，同じ場合は発話権保持 (hold) を予測結果とする．

3. マルチタスク学習による統合モデル

本研究ではベースラインモデルを，ターンテイキングの予測だけでなく，相槌，フィラーの生起も予測するマルチタスク学習による統合モデルへと拡張する．図 2 に本研究で提案するマルチタスク学習のモデルを示す．各層の括弧内はユニット数を表す．3 タスクで共通な層はタスク間で共通する特徴が，タスクごとに分かれている層は各タスク特有の特徴が，それぞれ学習されると考えられる．これより，相槌・フィラーとの関係性が考慮でき，ターンテイキングの予測が困難だった部分についてもより正確に予測できるようになる．入力特徴量はベースラインモデルと同様の 18 次元である．

ターンテイキングの学習・予測については，ベースラインモデルと同様に行う．相槌とフィラーについては，予測時点から将来 20 フレーム以内にそれぞれが生起する確率を出力する．そのため，正解ラベルは将来 20 フレーム以内に生起するか否かの 2 値である．また，学習時のパラメータ更新に用いる損失関数 $\mathcal{L}$ は以下の式で算出したものを使用する．学習時のパラメータ更新に用い

Smooth turn-taking using multitask learning with prediction of backchannels and fillers: *Kouhei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara (Kyoto Univ.)*

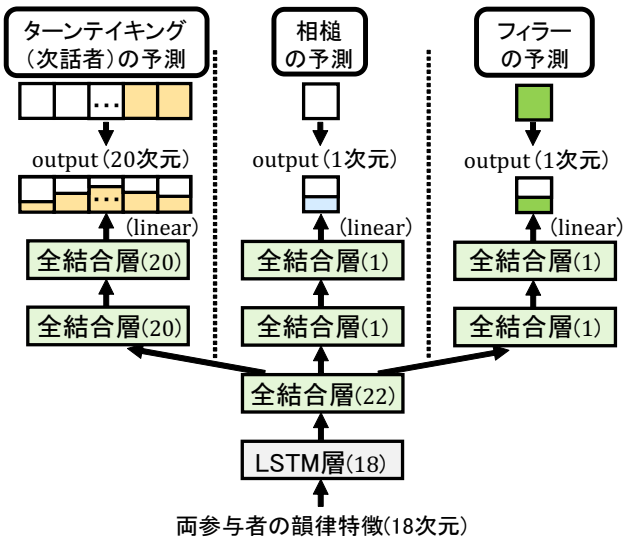

図 2: マルチタスク学習による統合モデル

る損失関数 $\mathcal{L}$ は以下の式で算出したものを使用する．

$$\mathcal{L} = \alpha \times \mathcal{L}_{ターン} + \beta \times \mathcal{L}_{相槌} + \gamma \times \mathcal{L}_{フィラー} \quad (1)$$

$\mathcal{L}_{ターン}$ は図 2 におけるターンテイキングの予測部分における損失関数，$\mathcal{L}_{相槌}$ は相槌の予測部分における損失関数，$\mathcal{L}_{フィラー}$ はフィラーの予測部分における損失関数であり，それぞれ平均二乗誤差を用いる．

4. 評価実験

モデルを評価するために，オペレータによって遠隔操作されたアンドロイド ERICA[6] と被験者との面接練習対話 16 セッション分の対話データを用いた．この対話では ERICA が面接官，被験者が面接受験者を演じ，アルバイトの面接という想定で行われた．16 セッションのうち，14 セッションを学習用，2 セッションをテスト用とする交差検定を行った．予測時点は全セッションで 2,194 箇所あり，そのうち，先行話者が発話権を保持した位置（hold）は 1,644 箇所，先行話者からもう一方の参与者へ発話権が移行した位置（take）は 550 箇所であった．

提案法であるマルチタスクモデル（3 節）とベースライン（シングルタスク）モデル（2 節）との性能比較を行った．これらのモデルは面接官（ERICA）と面接受験者（被験者）にそれぞれ用意されるが，面接官と面接受験者のモデルでは以下の設定は共通である．活性化関数は，出力層については線形関数を，それ以外の全結合層には ReLU を用いた．出力については [0，1] の値になるように，0 を下回る値は 0，1 を上回る値は 1 とした．パラメータの更新は，10 サンプル（60 sec の時系列データを 1 サンプルとした）をミニバッチとして，学習係数が 0.5×10^{-3} の RMSProp で行った．損失関数の重みは $\alpha = 1/3$，$\beta = 1/3$，$\gamma = 1/3$ とした．サンプル内の各時間フレームでの誤差は，正例と負例の数に大きな偏りが見られたため，タスク毎に式 (2) によって算出した．

$$\mathcal{L}_{frame,task} = \begin{cases} 負例比_{task} \times \mathcal{L}_{task,MSE}（正例） \\ 正例比_{task} \times \mathcal{L}_{task,MSE}（負例） \end{cases} \quad (2)$$

$\mathcal{L}_{task,MSE}$ は，ある時間フレームにおける各タスクの平均二乗誤差である．全セッションにおける正例と負例のそれぞれの数は，ターンテイキングは 59,812 と 127,562，相槌は 9,639 と 17,735，フィラーは 14,192 と 173,182 であった．エポック数は 2,400 とした．過学習を避けるために 正則化パラメータを 0.0012 とする $l2$ 正則化 を使用し，各層間に Dropout を適用した．ネットワークの実装には，TensorFlow 1.4.1 をバックエンドとする Keras 2.1.2 を用いた．評価尺度として適合率および再現率，およびそれらの調和平均である F 値を take と hold のそれぞれについて算出した．

表 1: 実験結果（take）

モデル	適合率	再現率	F 値
ベースライン	0.842	0.744	0.785
マルチタスク	**0.846**	**0.774**	**0.806**

表 2: 実験結果（hold）

モデル	適合率	再現率	F 値
ベースライン	0.917	**0.955**	0.935
マルチタスク	**0.926**	0.953	**0.939**

実験結果を表 1，表 2 に示す．hold については精度の差は見られなかったものの，take ではマルチタスク学習による効果が見られた．特に，take の再現率，hold の適合率が改善されている．前者は，発話権を獲得する意志を相手に示唆する役割のあるフィラーの予測が寄与していると考えられ，後者は，相手発話の継続を促す役割のある相槌の予測が寄与していると考えられる．以上より，ターンテイキングの学習時に相槌の予測とフィラーの予測の両方を考慮するマルチタスク学習が，ターンテイキングの予測精度向上に寄与することが確認できた．

5. おわりに

本稿では，マルチタスク学習の枠組みでターンテイキングと相槌，フィラーの予測を統一的に行うモデルを提案した．その結果，各タスク間の関係を考慮することで，ターンテイキングの予測精度が向上することを確認した．

今後は本研究で扱った面接練習対話以外の対話タスクにおける提案手法の有用性を検証していく．また，自律型アンドロイド ERICA のような実際の対話システムへの実装も行う．さらに，相槌・フィラー以外にも，視線やうなずきなどのターンテイキングに関係すると考えられる他のマルチモーダルな情報の利用も検討していく予定である．

謝辞 本研究は，JST ERATO 石黒共生ヒューマンロボットインタラクションプロジェクト JPMJER1401 の支援を受けて実施された．

参考文献

[1] John Kane *et al.*: “Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions,” *Interspeech*, pp.333–337, 2014.

[2] Gabriel Skantze: “Towards a General, Continuous Model of Turn-taking in Spoken Dialogue using LSTM Recurrent Neural Networks,” *SIGDIAL*, pp.220–230, 2017.

[3] Ryo Masumura *et al.*: “Online End-of-Turn Detection from Speech based on Stacked Time-Asynchronous Sequential Networks,” *Interspeech*, pp.1661–1665, 2017.

[4] 山口 貴史他: “傾聴対話システムのための言語情報と韻律情報に基づく多様な形態の相槌の生成,” 人工知能学会論文誌, 31 巻 (2016) 4 号 p. C-G31_1-10.

[5] 小磯 花絵. 話し言葉の書き起こし. 小磯 花絵（編）: “日本語コーパス 3: 話し言葉コーパス-設計と構築-,” pp.33–53, 朝倉書店, 東京, 2015.

[6] Koji Inoue *et al.*: “Talking with ERICA, an autonomous android,” *In Proc. SIGDIAL*, pp.212–215, 2016.