2T-04

移行適格場の予測に基づくターンテイキング予測

原 康平　　井上 昂治　　高梨 克也　　河原 達也

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

音声対話システムは様々な場面において実用化されている．自然で円滑な対話を実現するためには，ユーザの発話末における次話者の予測（ターンテイキング予測）が重要である．ターンテイキング予測のモデルとして，Long short term memory (LSTM) などの時系列ニューラルネットワークが用いられている [1, 2, 3]．また，特徴量には先行発話の韻律や言語の情報が用いられている．

本研究では，ターンテイキング予測において，移行適格場（TRP: Transition relevance place） の情報を利用することを提案する．TRP とは，ターンが交替しうる時点を指す [4]．人間どうしの自然な対話では，ターテイキングの判断には任意性が高いことがしばしばである．一方，TRP に関しては，ターンテイキングの判断に比べて任意性が低いといえる．TRP とターンテイキングには図 1 に示す二段階の関係があると考えらえる．はじめに，現在の発話の末尾が TRP であるか否かが判断される．その後，TRP である場合には，ターンが保持されるか獲得するかの判断が行われる．従来の方法では，TRP の情報を考慮せずに，任意性を含むターンテイキングのふるまいを直接予測していた．提案手法は，図 1 のように TRP とターンテイキングの予測を区別して行い，それらの予測結果を統合することで，ターンテイキングの予測精度向上を目指す．

2. 移行適格場（TRP）のアノテーション

本研究では，オペレータによって遠隔操作されたアンドロイド ERICA[5] と被験者との対話データを用いた．ERICA の音声は，オペレータが話した音声をリアルタイムで再生したものである．対話のタスクは，就職面接（13 対話），傾聴（13 対話），お見合い練習（33 対話）である．このデータに対して，TRP のラベルを人手により付与した．ここでは，長い発話単位 [6] の末尾において，聞き手の視点に立ち，ターンを取得してはならない箇所は非 TRP，取得しようとすれば可能である箇所を TRP とした．判断基準の一部として，隣接ペアおよび対話行為の情報を用いた．

例えば，質問に対する回答において，回答がまだ不十分な場合にはその箇所では非 TRP となる．ターンテイキングおよび TRP のラベルの内訳を表 1 に示す．ただし，単位は 200 ミリ秒を基準とする間休止単位（IPU: Inter-pausal unit）である．いずれの対話タスクでも，ターンテイキングのラベルは継続のほうが多いが，TRP の箇所に関しては，交替のほうが多くなっていることから，TRP と話者交替がある程度相関していることがわかる．

3. 提案モデル

提案モデルは，図 1 で示した仮定に基づいており，2 つのモデルそれぞれで学習を行う．1 つ目は，先行発話末が移行適格場（TRP）であるか否かを予測するものである．2 つ目は，TRP の箇所において話者交替または継続（ターンテイキング）を予測するものである．これらのモデルの出力を用いて，ターンテイキングの予測を行う．ただし，TRP 予測とターンテイキング予測はともに IPU 末で行う．

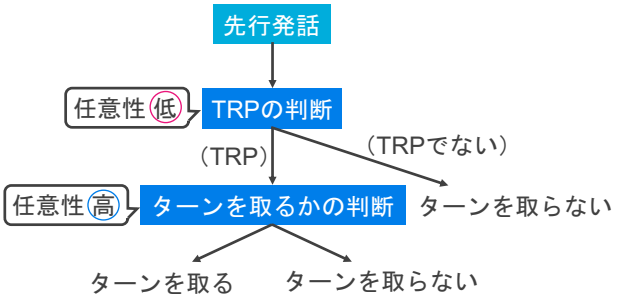

図 1: TRP 予測とターンテイキング予測の関係

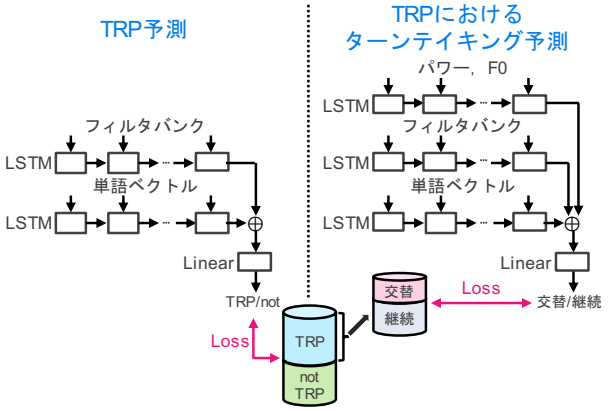

図 2: TRP 予測モデルおよび TRP におけるターンテイキング予測モデルの学習

3.1 TRP の予測

各 IPU の末尾において TRP であるか否かを予測する．図 2 の左側に学習の概要を示す．モデルは先行研究 [3] で用いられた LSTM に基づくものである．入力は，40 次元の対数メルフィルタバンクと 100 次元の単語ベクトルである．ただし，単語ベクトルは Word2Vec モデルを用いて抽出する．教師データは，人手でアノテーションされた TRP のラベルである．

3.2 TRP におけるターンテイキング予測

IPU 末が TRP である箇所についてターンテイキングを予測する．図 2 の右側に学習の概要を示す．モデルは TRP 予測の場合と同様の LSTM に基づくものである．特徴量としては，対数メルフィルタバンクと単語ベクトルに加えて，韻律特徴としてパワーと基本周波数（それぞれ一次と二次の変量を含む）の 6 次元を用いる．正解ラベルは，人手でアノテーションされたターンテイキング

Prediction of Turn-taking based on Prediction of Transition Relevance Place: *Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara (Kyoto Univ.)*

表 1: TRP とターンテイキングのラベルの内訳（括弧内は TRP の箇所における内訳）

タスク	TRP 予測		ターンテイキング予測	
	TRP	非 TRP	交替	継続
面接	529	1,860	446（ 440）	1,943（ 89）
傾聴	1,330	2,923	947（ 808）	3,306（ 522）
お見合い練習	4,583	5,415	3,551（3,223）	6,447（1,360）

表 2: ターンテイキングの予測結果

タスク	モデル	正解率	適合率	再現率	F 値	F 値マクロ
面接	ベースライン	91.2	74.6	66.2	70.2	82.5
	提案	92.3	71.1	85.0	77.4	86.4
傾聴	ベースライン	81.4	34.5	7.1	11.7	50.7
	提案	81.8	47.6	41.4	44.3	66.7
お見合い練習	ベースライン	77.6	68.4	58.8	63.2	73.5
	提案	76.4	65.3	59.7	62.4	72.6

のラベルである．ただし，TRP の箇所のみを用いる．

3.3 TRP 予測に基づくターンテイキング予測

3.1 節と 3.2 節のモデルを用いて，ターンテイキングを予測する．IPU 末における予測手順を以下に示す．

1. 3.1 節のモデルを用いて，TRP である確率 P_{TRP} を算出する．
2. 3.2 節のモデルを用いて，TRP において話者交替である確率 P_{Take} を算出する．
3. $P_{TRP} \times P_{Take} > 0.5$ ならば話者交替，そうでなければ話者継続を予測結果とする．

3. の条件式は，図 1 で示したように，TRP でない箇所では話者継続であるという前提に基づいている．これは，前述の対話コーパスでも実際に観測されている．

4. 評価

提案モデルの有効性を評価するために，TRP の情報を用いないベースラインモデルとの比較を行った．

4.1 条件

前述のアンドロイド ERICA と被験者との対話データを使用した．TRP のラベルは任意性が少なく対話タスクに依存しないと考えられるため，提案モデルの前段である TRP の予測モデルは全タスクのデータを用いて学習を行った．一方，後段のターンテイキングの予測モデルは，ターンテイキングのラベルには任意性が含まれ，その結果は対話タスクに依存すると考えられるため，モデルの学習は対話タスク毎に行った．上記の学習および評価は，5 分割交差検定により行った．ベースラインモデルは，TRP の情報を考慮せずに直接ターンテイキングを予測するものである．具体的には，提案モデルにおける後段の LSTM と同じで，学習には，TRP でない箇所も含むターンテイキングのラベルを用いる．評価指標として，正解率，話者交替の適合率・再現率・F 値，交替と継続の F 値の平均であるマクロ平均を用いた．

4.2 結果

ターンテイキングの予測結果を表 2 に示す．面接と傾聴では，F 値およびそのマクロ平均から，提案モデルによる精度向上が確認された．特に再現率が大きく改善されている．今回の対話データは，1 つのターンに複数の IPU が含まれる自然な対話であるため，ターンテイキングのラベルに関しては，継続の割合が大きい．そのため，従来の TRP を考慮しない場合では，少数ラベルである交替が出力されづらい傾向にあった．しかし，提案モデルでは，TRP の箇所を絞り込むことで，後段のターンテイキングの予測モデルの学習では，TRP ではなくほぼ継続である箇所を除くことができている．そのため，後段のターンテイキングの予測モデルの精度が向上したと考えられる．ただし，お見合い練習では精度の改善がみられなかった．お見合い練習では，他の 2 つの対話タスクに比べて，頻繁にターンが交替する．そのため，ターンテイキングのラベルにおいて，交替の割合がもともと高く，ベースラインモデルで十分な学習が行われたと考えられる．ただし，精度自体には改善の余地があるため，特徴量の追加を検討している．

5. おわりに

本稿では，移行適格場（TRP）の情報を用いたターンテイキングの予測手法を提案した．二段階で表される TRP とターンテイキングの関係性に基づき，任意性の低い TRP 予測の結果と，TRP におけるターンテイキング予測の結果を統合した．今後の課題として，特徴量の追加を検討している．また，TRP のラベルのアノテーションは，ターンテイキングのラベルの場合に比べて，作業コストが大きいため，学習済みの TRP 予測モデルを用いて自動アノテーションを行うデータ拡張も検討している．

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した．

参考文献

[1] Matthew Roddy *et al.* Investigating speech features for continuous turn-taking prediction using lstms. *INTERSPEECH*, 586–590, 2018.
[2] Ryo Masumura *et al.* Neural dialogue context online end-of-turn detection. *SIGDIAL*, 224–228, 2018.
[3] Divesh Lala *et al.* Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios. *ICMI*, 78–86, 2018.
[4] Sacks Harvey *et al.* A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.
[5] Tatsuya Kawahara. Spoken dialogue system for a human-like conversational robot ERICA. *IWSDS*, 2018.
[6] Yasuharu Den *et al.* Two-level annotation of utterance-units in japanese dialogs: An empirically emerged scheme. *LREC*, 2010.