2R-04

大規模テキストデータを用いた事前学習による音声対話の相槌予測

長　連成[1]　越智　景子[2]　井上　昂治[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

近年，スマートスピーカーなどが一般的な家庭にも置かれるようになり，音声対話システムは我々の生活に必要なものになりつつある．今後より社会的で長い対話を扱えるようになるためには，聞き手としての相槌が重要な役割を果たすと考えられる．人間どうしの対話において，メイナード [1] は相槌の果たす役割について「続けてというシグナル（continuer）」，「内容理解を示す表現」,「話し手の判断を支持する表現」,「相手の意見、考え方に賛成の意志表示をする表現」,「感情を強く出す表現」,「情報の追加、訂正、要求などをする表現」の６つを挙げている．我々はロボットを用いた音声対話の研究を進めており，ここでも相槌は同様の役割を果たすと考えられる．これまでに，相槌のタイミングやその種類を予測させるための研究を進めてきた [2].

上記の従来研究ではロジスティック回帰などの比較的単純なモデルが用いられている．本研究では，ユーザの発話の言語情報を入力として，その後に相槌をうつか否か，またうつ場合にはどの種類の相槌をうつべきかを予測するニューラルネットワークを構成する．しかし，音声対話における相槌のデータに関して，ニューラルネットワークを学習するための十分なデータ量を確保することは難しい．そこで，対話形式ではない大規模なテキストデータに対して，疑似的な相槌ラベルを付与することで，相槌予測モデルの事前学習を提案する．

2. 相槌予測モデル

予測モデルには，事前学習済みの BERT*を用いる．CLS トークンに対応する最終層の出力に線形層を追加する．入力は直前のユーザ発話の単語系列（最大 512 文字）である．出力はシステムの相槌のタイミングおよび種類に対応しており，ここでは「相槌なし」,「応答系」,「感情表出系」の３つである.「応答系」は「はい」や「うん」など,「感情表出系」は「へー」や「おー」などの驚きや関心を表すものにそれぞれ対応する．

3. テキストデータを用いた事前学習

後述する音声対話データでは十分な学習データを確保することが難しいため，対話形式ではないテキストデータを用いて疑似的な相槌ラベルを生成し，追加の事前学習を行う．使用するテキストデータには，できるだけ話し言葉に近いもの用いるために，日本語書き言葉コーパス（BCCWJ）の「国会議事録」を用いる．このデータのうち９割は事前学習に用いるが，残りの１割は事前学習の効果を評価する際のテストデータに使用した．

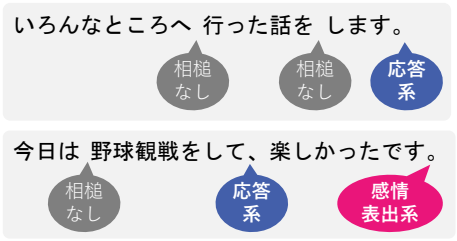

図 1: 事前学習のための疑似的な相槌ラベルの作成例

疑似的な相槌ラベルの付与方法について述べる．テキストデータには句読点が付与されているためこの情報を活用する．図 1 に例を示す．まず，句読点が付与されている箇所では相槌がうたれると仮定する．相槌がうたれない箇所については，句読点が付与されていない箇所でさらに助詞の部分を採用する．ただし，助詞の箇所は句読点のそれより多く，データ不均衡が生じるため，すべての助詞を採用するのではなく，ランダムにサブサンプリングする．句読点が付与されている箇所では，さらに応答系と感情表出系のどちらか相槌ラベルを付与する．その基準として，直前の文脈の感情極性の情報を活用する．辞書ベースの感情極性判別器 Oseti†を用い，極性が存在する場合には感情表出系，存在しない場合には応答系とした．

以上により作成した疑似的な相槌ラベルを用いて，事前学習済みの BERT に対して追加の事前学習を行う．学習データ量は，相槌なしが 4,220 件，応答系が 1,755 件，感情表出系が 6,562 件となった．

4. 音声対話データでのファインチューニング

最後に，音声対話コーパスにおける実際の相槌のデータを用いて相槌予測モデルをファインチューニングする．このデータには，アンドロイド ERICA を用いた音声対話コーパス [3] を用いた．この対話は遠隔操作された ERICA と被験者との一対一の対話である．ERICA の操作は別室のオペレータによって行われており，オペレータが話した音声をそのまま ERICA のスピーカから再生している．コーパス全体を通して，オペレータは４人である．このコーパスにはいくつかの対話タスクが設定されているが，ここではお見合い対話のデータを使用する．対話数は 84 で，ERICA の役割はお見合いの練習相手である．図 2 にオペレータの相槌の分布を示す．応

Prediction of Backchannels in Spoken Dilogue by Pre-training with Large-scale Text Data: *Rensei Cho, Keiko Ochi, Koji Inoue, Tatsuya Kawahara (Kyoto Univ.)*

*https://github.com/cl-tohoku/bert-japanese

†https://github.com/ikegami-yukino/oseti

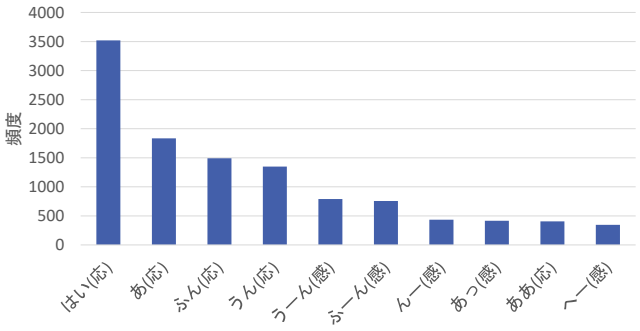

図 2: お見合い対話におけるオペレーターの相槌の種類の分布（頻出上位 10 件,「応」は応答系,「感」は感情表出系の相槌をそれぞれ表す）

表 1: 相槌予測の結果（F 値 [%]）

	ベースライン	提案手法
相槌なし	64.2	64.3
応答系	49.9	53.6
感情表出系	0.7	8.1

答系も感情表出系も一定数以上出現していることがわかる．今回はこの相槌のデータを用いてその予測を試みる．ただし予測を行う単位は，200 ミリ秒のポーズを基準とする間休止単位（IPU）とする．最終的には，合計サンプル数が 15,671 で，このうち相槌なしが 7,665，応答系が 5,865，感情表出系が 2,141 となった．上記のデータのうち 9 割を学習用，1 割をテスト用とした．

5. 評価

テキストデータによる事前学習の効果を検証した．ベースラインはテキストデータによる追加の事前学習を行わず，事前学習済み BERT モデルを用いて音声対話データによるファインチューニングをそのまま行ったものである．評価指標は，相槌なし，応答系，感情表出系のそれぞれの F 値である．

結果を表 1 に示す．テキストデータによる事前学習を行った方が，応答系および感情表出系の F 値が向上した．ただし，感情表出系の精度については絶対的な数値は低いままである．

表 2 と表 3 にベースラインと提案手法のそれぞれにおける予測の混合行列を示す．これらを比べると，提案手法では感情表出系が一定程度出力されていることがわかり，感情極性の情報を用いたテキストデータによる事前学習の効果が伺える．ただし，両者とも感情表出系のところを応答系と出力した箇所（151 および 168 サンプル）が多いことからも，応答系と感情表出系の区別が依然として難しいこともわかる．

最後に実用的な観点として，相槌のタイミングのみを予測した場合の評価も行った．ここでは応答系と感情表出系の種類の区別はせずに，つまり同じ予測モデルの出力を 3 次元から 2 次元（相槌ありとなし）へまとめた上で評価を行った．その結果，テキストデータによる事前学習を行わなかった場合（ベースライン）には相槌ありの F 値が 59.5%，行った場合（提案手法）には F 値が 66.6%となった．このことからもテキストデータによる事前学習の効果が分かる．

表 2: ベースラインによる予測の混同行列

		予測		
		相槌なし	応答系	感情表出系
正解	相槌なし	644	210	0
	応答系	383	372	3
	感情表出系	124	151	1

表 3: 提案手法による予測の混同行列

		予測		
		相槌なし	応答系	感情表出系
正解	相槌なし	587	250	17
	応答系	291	431	36
	感情表出系	94	168	14

6. おわりに

本稿では，音声対話における相槌のタイミングと種類の予測において，テキストデータに対する疑似的な相槌ラベル付与による事前学習を提案した．音声対話コーパスによるファインチューニングおよび評価の結果，相槌のタイミングおよび種類の予測において事前学習の効果がみられた．ただし，感情表出系の予測については精度自体が十分とはいえないため，疑似ラベルの作成において極性以外の情報を検討してく必要があるといえる．また，言語情報だけでなく，韻律情報を用いることも今後の課題である．

謝辞

本研究は科研費（19H05691）の支援を受けた．

参考文献

[1] 泉子・K・メイナード, 会話分析，くろしお出版，1993.

[2] 山口貴史, 井上昂治, 吉野幸一郎, 高梨克也, Nigel G. Ward, 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, 2016.

[3] T. Kawahara, Spoken Dialogue System for a Human-like Conversational Robot ERICA, *IWSDS*, 2018.

[4] Y. Den, N. Yoshida, K. Takanashi, H. Koiso, Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations, *Oriental COCOSDA*, 2011.