系列変換モデルに基づく傾聴的な応答表現の生成

村田 匡輝[†]　大野 誠寛[‡]　松原 茂樹[††]
[†]豊田工業高等専門学校 情報工学科　[‡]東京電機大学 未来科学部
[††]名古屋大学 情報連携統括本部
murata@toyota-ct.ac.jp

1 はじめに

語ることは人間に備わる基本的な欲求である．語るという行為は，聞き手がいて初めて成立する．社会の個人化が進み，聞き手不在の生活シーンが増加している．人が語れる機会を増やすことは現代社会の重要な課題といえる．

上述の課題に対して，コミュニケーションロボットやスマートスピーカなどの情報機器が語りを聞く役割を担うことが考えられる．機器が聞き手として認められるには，

1. 語りを傾聴する機能
2. 語りを傾聴していることを話し手に伝達する機能

を備える必要がある．1. は音声認識や言語理解の技術により実現される．一方，2. の明示的な手段は語りに応答することであり，具体的にはジェスチャや発話の表出が有力である．以下では，2. を達成する発話，すなわち，傾聴的な態度を示す目的で語りに応答する発話を**傾聴的応答**と呼ぶ．傾聴的応答の代表は相槌であり，その生成法が提案されているものの [1,2]，傾聴を示す応答は相槌以外にも存在する [3]．

本稿では，語りの聞き手を担う情報機器の実現に向け，語りに対して傾聴的応答を付与したデータ（以下，**応答データ**）を使用した深層学習による応答表現の生成手法について述べる．本手法では，ある時点までの語りの単語列を入力とし，Long Short-Term Memory (LSTM) を用いた系列変換モデルにより，応答表現を生成する．入力として与える単語列の単位，モデルを構成するパラメータ等を変化させた複数のモデルを学習し，傾聴的応答の特性から定めた評価指標により，各モデルを比較・評価する．応答データを用いた応答表現生成実験の結果，多様な応答表現を生成できていることを確認した．

2 傾聴的応答の生成

本研究では，傾聴的応答を適切に表出することにより，語りの傾聴を話し手に伝達する機能の実現を目指している．情報機器が応答を生成するにあたり，応答表現の選択と生成タイミングの決定という 2 点を解決する必要がある。生成の効果を高めるという観点から，

- 生成の頻度が高くかつ自然なタイミングであること
- 自然でかつ多様な応答表現であること

が要件となる．

本研究では，応答のタイミングは既知であることを前提に，ある時点から応答タイミングまでの語りを入力とし，自然かつ多様な応答表現の生成を目指す．なお，生成するのは応答表現の文字列のみであり，イントネーション，アクセント等の音響的特徴は考慮しない．実際には，文字列としては同一の「はい」であっても，語尾を上げるか下げるかによってその意味が異なるため，音響的特徴の考慮は今後の課題である．

関連として，聞き役対話システムの研究 [3–5] が挙げられる．これらは，システムが問いかけや話題提供によってユーザの話の展開に積極的に関わることによりユーザから話を引き出す技術の開発を主な目的としている．いずれも聞き手の行為として有効であるが，本研究が目指す傾聴的応答では，対話の発話権を保持する応答を行うことなく，あくまで一方的に応答を行いながらも，ユーザに話を聞いているということを伝えることにより，対話を促進することを目指している．そのため，目標とする対話の形態に違いがある．

深層学習による応答生成に関する研究がいくつか行われている [6–8]．これらの研究では，発話と応答のペアを大量に収集し，リカレントニューラルネットワーク (RNN) や LSTM を用いた系列変換モデルにより応答を生成する．本研究でも，LSTM を用いた系列変換による応答生成という手法を採用する．ただし，これらの研究では，ある程度ターンの境界が明確な対話を対象としている．一方で，本研究は，独話的に語られる発話に対して，発話権を保持しない応答を行う形態を対象としている．そのため，発話と応答のペアは明確ではない．そこで，入力として複数の種類の単語列を検討し，モデルの性能を比較する．

3 系列変換による応答生成

3.1 応答生成モデル

傾聴的応答の第一形態素の発話時間に基づき，語りにおける応答タイミングを定める．本研究では，応答の第一形態素の発話開始時間の直前に発話が開始された語りの形態素の発話終了時間を応答タイミングとした．図 1 に応答タイミングの例を示す．

本研究では，LSTM を使用した系列変換モデルにより，応答表現の生成を行う．ある時点から応答タイミングまでの語りの単語列と，応答の単語列のペアを用

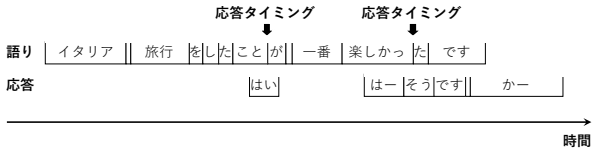

図 1: 応答タイミング

い，系列変換モデルを学習する．ある時点から応答タイミングまでの語りの単語列 $X = (x_1, ..., x_n)$ をエンコーダに入力し，その出力を，デコーダの入力へ接続することにより，応答の単語列 $Y = (y_1, ..., y_m)$ を獲得する．入力単語 x_i, y_j は単語埋め込み層によって，d 次元の埋め込みベクトルに変換する．エンコーダには双方向エンコーダを用いる．活性化関数には softmax を，誤差の計算には cross entropy を使用する．また，最適化手法には Adam を用いる．応答生成モデルの概要図を図 2 に示す．

応答表現の生成においては，デコーダに文末記号（<eos>）を入力し，確率が最大となる単語を順に出力する．文末記号が出力されたときに応答表現の生成を終了する．

3.2 モデルの学習・調整

応答生成モデルに入力する，ある傾聴的応答に対する語りとして，ある時点からその傾聴的応答がなされるタイミングまでの語りの単語列を与える．ただし，語りは独話的に発話されており，ある傾聴的応答に対応する語りの開始位置は明確ではない．そこで本研究では，以下の三種類の開始位置からの単語列を入力としたモデルを学習する．

(1) 直前の応答タイミングの直後

(2) 三つ前の応答タイミングの直後

(3) 直前の絶対境界・強境界 [9] の直後

また，モデル (2) では，語りの単語列中に必ず複数の応答タイミングが含まれる．語りと傾聴的応答の両方の文脈を考慮することを目的に，モデル (4) として，三つ前の応答タイミングからの語りの単語列に，各応答タイミングにおける応答の単語列を含めたものも学習する．

4 実験

各モデルの性能を確認するため，応答データを用いた応答生成実験を実施した．

4.1 実験概要

著者らはこれまでに，高齢者の語りに対して傾聴的応答を付与した応答データの収集を行ってきた [10]．応答データは，語りのデータとして，高齢者のナラティブコーパス JELiCo[1] を使用し，その音声に対して高度な接客スキルを要する業務経験を有する作業者 1 名が応答を実施することで作成した．図 3 にデータの例を示す．人が知覚できるポーズで分割した単位を発話

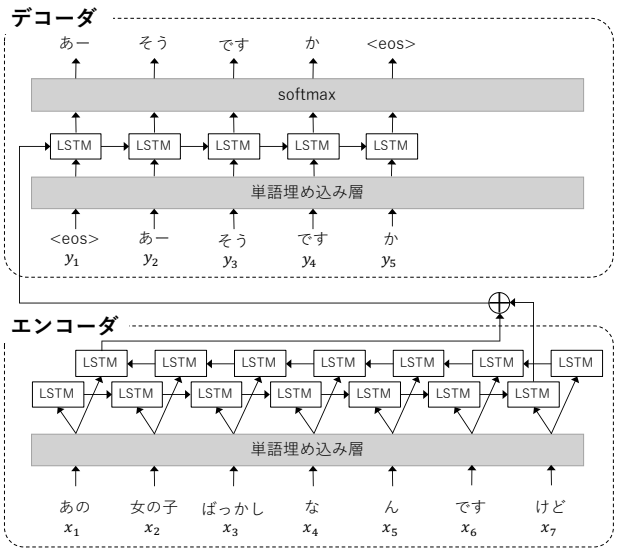

図 2: 応答生成モデル（隠れ層 1 層）

単位と定め，発話単位の形態素情報，及び，形態素単位の発話時間を付与している．実験データとしてこの応答データを用いる．

3.2 に示した各応答生成モデルを学習し，入力として与えられる語りの単語列に対し，各モデルを用いて応答表現の単語列を生成する．以下の値については，各モデルごとに，開発データを用いて，損失が最小になるようなものを探索する．探索回数は 100 回とした．

- エポック数
- ミニバッチサイズ
- エンコーダの隠れ層の数
- 埋め込みベクトルの次元数
- ドロップアウト率

モデルの学習には，応答データのうち，9,100 ペアを，モデルチューニング用の開発データには 3,153 ペアを用いた．また，今回の実験では，応答生成の入力として，チューニングに用いた開発データの 3,153 個の語りを使用した．語りの語彙数は 5,427，傾聴的応答の語彙数は 1,354 であった．

生成した応答表現と正解の傾聴的応答を比較し，評価を行う．評価指標として，応答データの傾聴的応答との一致率に加え，モデルを多面的に評価するため，以下の 3 種類を用いた．

- BLEU
- 言語モデル確率
- 応答表現の種類数（文字列の異なり数）

言語モデル確率によって生成した応答表現の応答らしさを，種類数によって多様性を評価する．言語モデル確率の計算には，応答データに含まれる全傾聴的応答の単語列から LSTM を用いて学習したニューラル言語モデルを使用した．

[1] http://sociocom.jp/software.html

高齢者の語り		傾聴的応答	
01:07:02 - 01:11:51	イタリア旅行をしたことが一番楽しかったです	01:08:99 - 01:09:25	はい
		01:09:87 - 01:11:57	はーそうですかー
		01:11:57 - 01:12:85	素敵ですねー
01:14:52 - 01:19:26	もう二度と行けないかなと思いながら行ってきましたけど	01:13:73 - 01:14:30	ふーん
		01:14:30 - 01:15:04	イタリア旅行
		01:16:54 - 01:18:04	いえいえそんなー
		01:18:52 - 01:19:76	あーそうですかー

図 3: 応答データの例

表 1: 実験結果

モデル	一致率 (%)	BLEU	言語モデル確率 (%)	種類数
(1) 直前の応答タイミング エポック数=5，ミニバッチサイズ=384，隠れ層=2 埋め込みベクトル次元数=896，ドロップアウト率=0.4	34.38	0.0098	93.14	30
(2) 三つ前の応答タイミング エポック数=5，ミニバッチサイズ=256，隠れ層=1 埋め込みベクトル次元数=640，ドロップアウト率=0.2	31.40	**0.0274**	88.37	**34**
(3) 直前の絶対境界・強境界 エポック数=5，ミニバッチサイズ=640，隠れ層=1 埋め込みベクトル次元数=960，ドロップアウト率=0.3	36.03	0.0178	**93.27**	20
(4) 三つ前の応答タイミング＋傾聴的応答 エポック数=5，ミニバッチサイズ=512，隠れ層=1 埋め込みベクトル次元数=576，ドロップアウト率=0.3	**36.79**	0.0267	91.23	22

表 2: エポック数を 20 としたときの 実験結果

モデル	一致率 (%)	BLEU	言語モデル確率 (%)	種類数
(1) 直前の応答タイミング	26.36	0.0153	83.30	**205**
(2) 三つ前の応答タイミング	27.81	0.0182	87.10	94
(3) 直前の絶対境界・強境界	28.70	0.0131	**90.09**	52
(4) 三つ前の応答タイミング＋傾聴的応答	**30.16**	**0.0221**	86.45	102

4.2 実験結果

実験結果を表 1 に示す．表 1 より，どのモデルにおいても，言語モデル確率は 88%以上となっており，傾聴的応答らしいものが生成できていることが分かる．

また，一致率は 35% 前後となっているが，正解と一致しているものの多くは「はい」であった．応答データにおいては，相槌の「はい」の出現頻度が最も大きく，全体の 4 割程度を占める．「はい」の次に，「えー」「えーえー」と続く．損失が最小になるように各モデルをチューニングした結果，「はい」や「えー」といった出現頻度の大きい応答表現が多く生成されるようになったと考えられ，生成された応答表現の種類数は 20 から 35 個であった．正解における傾聴的応答の種類数は 628 個であり，各モデルが生成した応答表現の多様性は十分ではない．

4.3 エポック数を増加させたモデルの学習

応答表現の多様性を高めるため，エポック数を 20 に増やし，モデルの学習，応答表現の生成を行った．エポック数以外のパラメータは変化させない．結果を表 2 に示す．損失が大きくなることが確認され，一致率や言語モデル確率に減少が見られるものの，生成された応答表現の種類数は 50 から 200 個程度に増加した．

以下に，生成された応答表現の例を示す[2]．なお，応答の括弧中の数字は順に，使用したモデルとエポック数を表す．

正解と一致

以下は，エポック数を増やしたことにより，正解と一致する応答表現が生成できた例である．

語り：/んで今度の総会にね顔をだすと/うー約束を/
　　やりまして/
　正解：あーそーなんですねー
　応答((2), 5)：はいー
　応答((2), 20)：あーそーなんですねー
語り：/えー踊ったり/歌ったり/またあー/
　正解：ふーん
　応答((2), 5)：あっ
　応答((2), 20)：ふーん

[2] "/" は応答タイミングを示す．また，いずれのモデルで生成された応答表現の例においても，文脈の確認のため，三つ前の応答タイミングの直後からの語りの単語列を示している．

正解と意味的に類似

下記の例のように，正解とは一致しないものの，意味合いとしては類似した応答表現が生成できているものも存在した．

語り：(/けど/ご近所だったので) /本当に/
正解：はいびっくりですねー大変でした
応答((1), 5)：えー
応答((1), 20)：サプライズ

語り：/そんなようなことで/にさんち前に帰って来たばっかり/なので変わった/
正解：お疲れですねー
応答((2), 5)：あー
応答((2), 20)：大変

正解と不一致だが適切（称賛）

下記は，語りの内容に対して称賛を送る役割の応答表現の例である．称賛を送ることは，話者の語りを促すうえで有効な方法の一つである．

語り：/思っています/これからも継続して/そこの会社にはえー続けていきたいなと/
正解：あっ
応答((2), 5)：そーですねー
応答((2), 20)：素敵ですねー

語り：/重量んーザックに背負ってる重量/が約 30 キロぐらい/というもので歩いたことがあります/
正解：一日
応答((4), 5)：あーそーなんですねー
応答((4), 20)：すごいですねー

正解と不一致だが適切（語の繰り返し）

傾聴的応答の一つとして，相手発話中の語を繰り返して応答する繰り返し応答が存在する．本手法により，わずかではあるが繰り返し応答の生成が確認できた．

語り：/もーしたいことって/いうのはもーあれですね特に/もーないですねー/
正解：そですかーございませんかぁ
応答((2), 5)：えーえー
応答((2), 20)：ないですねー

語り：(/通っ/て) /新宿まで出ました/
正解：はー
応答((1), 5)：あっそーなんですねー
応答((1), 20)：あっはい新宿まで

以上の例のように，エポック数を増やすことにより，多様な応答表現が生成できていることを確認した．しかし，不適切な応答表現も多く生成されているため，多様な応答表現を生成しつつ適切さを確保することが今後の課題となる．

5 おわりに

本稿では傾聴性を備えた情報機器の実現を目的とした応答表現の生成手法を提案した．本手法では，高齢者の語りの単語列を入力とし，LSTM を用いた系列変換モデルにより，応答表現の単語列を生成する．

応答生成実験を実施した結果，傾聴的応答らしい応答表現を生成できていることが確認できた．また，学習時のエポック数を増加させることにより，称賛を送る応答，繰り返し応答等の生成が実現できた．

今後の課題として．応答データの拡充が挙げられる．今回使用した学習データは 9,100 ペアであり，損失を最小にするようモデルをチューニングした結果，学習データ中の出現頻度の大きい応答表現が多く生成された．エポック数を増加させることで，応答表現の多様性は増加するが，不適切な生成結果も存在した．データ量を増加させることにより，この問題を解決できる可能性がある．また，本研究では応答のタイミングを既知と仮定したが，実際に情報機器を実現する際には，応答タイミングの推定手法も必要となる．

謝辞　高齢者のナラティブコーパスは，奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室から提供いただいた．本研究は，一部，科学研究費補助金（挑戦的萌芽研究）(No. 15K12095) により実施したものである．

参考文献

[1] 大野誠寛, 神谷優貴, 松原茂樹. 対話コーパスを用いた相づち生成タイミングの検出. 電子情報通信学会論文誌, Vol. J100-A, No. 1, pp. 53–65, 2017.

[2] 山口貴史, 井上昂治, 吉野幸一郎, 高梨克也, Nigel G. Ward, 河原達也. 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, Vol. 31, No. 4, pp. C–G31_1–10, 2016.

[3] 下岡和也, 徳久良子, 吉村貴克, 星野博之, 渡部生聖. 音声対話ロボットのための傾聴システムの開発. 情報処理学会論文誌, Vol. 53, No. 12, pp. 2787–2801, 2017.

[4] 小林優佳, 山本大介, 土井美和子. 高齢者対話インタフェース —発話間の共起性を利用した傾聴対話の基礎検討—. FIT2011 講演論文集, Vol. 10, No. 2, pp. 253–256, 2011.

[5] 目黒豊美, 東中竜一郎, 堂坂浩二, 南泰浩. 聞き役対話の分析および分析に基づいた対話制御部の構築. 情報処理学会論文誌, Vol. 53, No. 12, pp. 2787–2801, 2012.

[6] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In *Proc. of ACL-IJCNLP 2015*, pp. 1577–1586, 2015.

[7] Oriol Vinyals and Quoc V. Le. A neural conversational model. In *Proc. of ICML 2015 Deep Learning Workshop*, 2015.

[8] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. Hierarchical neural network generative models for movie dialogues. In *Proc. of AAAI*, pp. 3776–3783, 2016.

[9] 高梨克也, 丸山岳彦, 内元清貴, 井佐原均. 話し言葉の文境界 — CSJ コーパスにおける文境界の定義と半自動認定—. 言語処理学会第 9 回年次大会発表論文集, pp. 521–524, 2003.

[10] 村田匡輝, 大野誠寛, 松原茂樹. 話し手の語りに傾聴的な応答の収集. 言語処理学会第 23 回年次大会発表論文集, pp. 636–638, 2017.