ユーザーの発話意図理解に基づくインタビュー発話の生成*

☆松好祐紀, 滝口哲也, 有木康雄 (神戸大)

1　はじめに

現在，日本では超高齢化社会に対応するための方針の一つとして，高齢者の健康維持や介護予防，孤立防止などにつなげるため，退職後の社会参加を促している．しかし，実際に社会参加を行っている高齢者の割合は 2016 年の時点で 6 割程度である．

本研究では，高齢者を対象に，「社会参加をどのように考えているか」について聞き取るインタビュアーエージェントの構築を目標としている．本稿では，インタビュアーエージェントの中核である，言語理解部，対話管理部，言語生成部に関して述べる．言語理解部では，LSTM Encoder-Decoder を用いてユーザーの発話意図をフレーム，発話文中に現れるキーワードをスロットとして推定する．また，言語生成部では，言語理解部と同じく LSTM Encoder-Decoder を用いて，フレーム，スロット形式で表されたシステムの発話意図から，システムの発話を生成する．

本稿では，言語理解部，言語生成部に関しては，これまでに研究してきた「オセロゲーム中にユーザーを支援する質問応答システム」[1] において実装したモデルを用いて説明する．現在，高齢者を対象としたインタビューのデータを収集している．しかし，質問応答においても，インタビュー対話においても，ユーザーの発話意図と発話中のキーワードを理解し，それらを基にシステムの発話意図の決定，発話生成を行うことは共通していると考えているので，[1] の研究で構築したモデルを今回の研究に流用する．

2　これまでの研究

これまでの研究では，ユーザーがオセロゲームを行う際に生じる質問に，ゲームシステム自体が対話的に回答をすることで，ユーザーの理解を促進させることを目標としていた．想定していたシステムの構成を，Fig. 1 に示す．まず，ユーザーの質問が音声認識され，文字列の形で言語理解部に渡される．そして，言語理解部で推定されたユーザーの発話意図を，対話管理部においてシステムの発話意図に変換し，最後に言語生成部において，システムの発話意図からシステム発話を生成する，という形を想定している．以降，このシステムにおける言語理解部，言語生成部の提案モデルについて述べる．

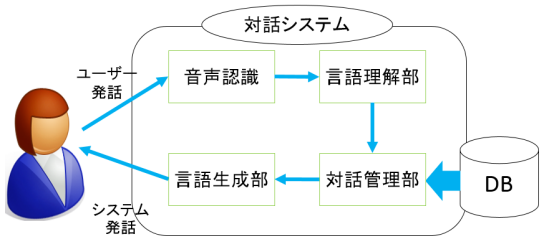

Fig. 1　システム構成図

3　提案システム

3.1　言語理解部

言語理解部では，ユーザー発話のフレーム，スロットを推定する．これまでの研究 [1] では，フレームはユーザーの大まかな意図を表す「質問タイプ」として定義している．また，スロットは，質問文中のキーワード（以下，「質問キーワード」と呼ぶ）の種類を「質問キーワードクラス」として定義している．例えば，質問キーワードが「X 打ち」なら，質問キーワードクラスは「オセロ用語」になる．質問タイプに関しては，計 15 種類を定義した．以下に例を示す．

- 理由: システムの回答などに対して生じる疑問に関する質問
- 場所: 盤面上の場所、座標に関する質問
- 結果: 指定した座標に打った場合の展開に関する質問
- 勝敗: 勝敗や形勢に関する質問

また，質問キーワードクラスに関しては，計 13 種類を定義した. 以下に例を示す.

- 用語: X 打ち、開放度など、オセロの専門用語
- 座標: b8、f7 など、オセロ盤面のマスの呼び方

3.2　フレーム，スロットの推定

提案モデルは Fig. 2 のような，LSTM Encoder-Decoder で実装した (以下，$Model_{SLU}$ とする)．質問文を形態素に分解し，それぞれを one-hot な単語ベクトルに変換して $(x_1 \dots x_m)$word embedding を行い，モデルに入力する．文末記号 <eos> が入力された時点の隠れ層 h_{qtype} から質問タイプの推定値を計算する．そして，<go> が入力された後，質問キーワードクラス $(y_1, y_2 \dots, y_j)$ が順番に生成される [2]．このモデルでは，Attention 機構を導入している [3]．各デ

*Generation of interview utterance based on user's utterance intention understanding. by MATSUYOSHI, Yûki, TAKIGUCHI, Tetsuya, ARIKI, Yasuo (Kobe University)

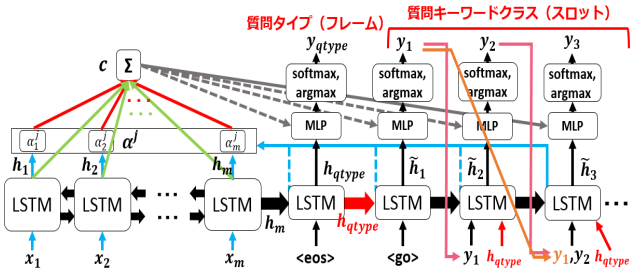

Fig. 2　$Model_{SLU}$ の概略図

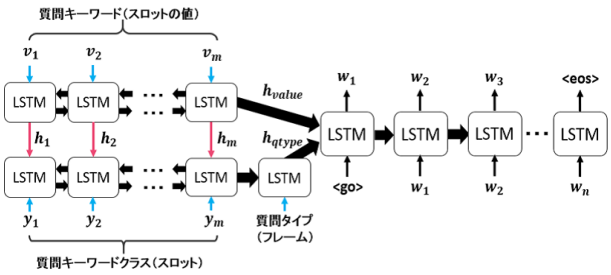

Fig. 3　$Model_{SLG}$ の概略図

コードステップの中間層 $\boldsymbol{h_{qtype}}$，$\widetilde{\boldsymbol{h}}_{\boldsymbol{j}}$(j=1,2,...,m) と，質問文入力時に保持しておいた，エンコーダの中間層の出力 $\boldsymbol{h_i}$(i=1,2,...,m) を利用して、Fig.2 の α_i^{qtype}，α_i^j を計算する．j はデコードステップを表す．

$$\alpha_i^{qtype} = \frac{\exp(\boldsymbol{W_1}^T tanh(\boldsymbol{W_2 h_i} + \boldsymbol{W_3 h_{qtype}}))}{\sum_{n=1}^{m} \exp(\boldsymbol{W_1}^T tanh(\boldsymbol{W_2 h_n} + \boldsymbol{W_3 h_{qtype}}))}$$

$$\alpha_i^j = \frac{\exp(\boldsymbol{W_1}^T tanh(\boldsymbol{W_2 h_i} + \boldsymbol{W_3}\widetilde{\boldsymbol{h}}_{\boldsymbol{j}}))}{\sum_{n=1}^{m} \exp(\boldsymbol{W_1}^T tanh(\boldsymbol{W_2 h_n} + \boldsymbol{W_3}\widetilde{\boldsymbol{h}}_{\boldsymbol{j}}))} \quad (1)$$

$\boldsymbol{W_1}$，$\boldsymbol{W_2}$，$\boldsymbol{W_3}$ は学習するパラメータである．$\alpha(i)$ を $\boldsymbol{h_i}$ の重みとし，コンテキストベクトル $\boldsymbol{c^{qtype}}$, $\boldsymbol{c^j}$ を計算する．

$$\boldsymbol{c^{qtype}} = \sum_{i=1}^{m} \alpha_i^{qtype} \boldsymbol{h_{qtype}}, \quad \boldsymbol{c^j} = \sum_{i=1}^{m} \alpha_i^j \widetilde{\boldsymbol{h}}_{\boldsymbol{j}} \quad (2)$$

$\boldsymbol{c^{qtype}}$, $\boldsymbol{c^j}$ と $\boldsymbol{h_{qtype}}$, $\widetilde{\boldsymbol{h}}_{\boldsymbol{j}}$ から，以下の式により y_{qtype}，y_j を計算する．$\boldsymbol{W_4}$，$\boldsymbol{W_5}$ は学習するパラメータである．

$$y_{qtype} = argmax(softmax(tanh(\boldsymbol{W_4 c} + \boldsymbol{W_5 h_{qtype}})))$$

$$y_j = argmax(softmax(tanh(\boldsymbol{W_4 c} + \boldsymbol{W_5}\widetilde{\boldsymbol{h}}_{\boldsymbol{j}}))) \quad (3)$$

$Model_{SLU}$ の特徴としては，デコーダでの各時間ステップでの入力に h_{qtype} を加えている．また，現在の時間ステップ j の出力 y_j の計算に，2 つ前の時間ステップ $j-2$ の隠れ層の出力 y_{j-2} も利用する．

3.3　言語生成部

言語生成部のモデルでは，システムの発話意図からシステムの発話を生成する．モデルは Fig. 3 のような Encoder-Decoder で構築している ($Model_{SLG}$)．エンコーダは，発話意図のスロット $(y_1 \dots y_m)$ とフレームをエンコードする部分と，スロットの値 $(v_1 \dots v_m)$ をエンコードする部分に分かれる．デコーダでは，<go> が入力された後，システム発話の形態素 $(w_1, w_2 \dots, w_n)$ が順番に出力される．質問キーワードクラス（スロット）の各エンコードステップで，質問キーワード（スロットの値）の各エンコードステップの隠れ状態を入力している理由は，両者のエンコードのアライメントをとるためである．デコーダへ入力される隠れ状態は，学習パラメータ $\boldsymbol{W_6}$，$\boldsymbol{W_7}$ を用いて，$\boldsymbol{h_{value}}$ と $\boldsymbol{h_{qtype}}$ を連結したベクトルである．

$$\boldsymbol{W_6 h_{value}} \oplus \boldsymbol{W_7 h_{qtype}}$$

Table 1　joint モデル推定の質問タイプ推定に関する結果．（推定率の単位：%）

モデル	質問タイプ			質問キーワードクラスの推定率
	推定率	再現率	適合率	
$Model_{QT}$	85.3	0.81	0.83	-
$Model_{QK}$	-	-	-	81.4
$Model_{SLU}$	85.3	0.79	0.83	83.6

4　実験

4.1　言語理解部の実験概要

データセットは，実際にオセロゲームをプレイする際に生じるユーザーの質問文データ集合 1788 文であり，各文に対して，教師データとして，質問タイプ，質問キーワードクラス，質問キーワードを人手でアノテーションした．1788 文の内，1605 文を学習データ，183 文をテストデータとして使用した．テストでは，テストデータを入力して，183 文の内，どれだけ正しく質問タイプ，質問キーワードクラスを推定出来たかを調べた．また，$Model_{SLU}$ のエンコーダ部分は共通で，質問タイプのみを推定する $Model_{QT}$，質問キーワードクラスのみを推定する $Model_{QK}$[1] も構築し，比較を行った．

LSTM では，word embeddings はランダムに初期化され，サイズは 250 次元である．ドロップアウトの確率は 0.1，モデルの最適化には Adam を使用した．Adam のパラメーは ($\alpha = 0.0001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$) となっている．

4.2　言語理解部の実験結果

実験結果を Table 1 に示す．質問キーワードクラスに関しては，$Model_{QK}$ に比べ $Model_{SLU}$ では推定率が向上したが，質問タイプに関しては，$Model_{QT}$ と比べ，推定率の向上は見られなかった．質問キーワードクラスは質問タイプに比べて推定が複雑になるので，推定の信頼度が低くなる．よって，両者を組み合わせて推定すると，質問タイプの推定に影響が出た

correct:[切り返し って どういう こと です か ?]
predicted:[切り返し って 何 か な ?]

correct:[山 を 作り たい の です が]
predicted:[山 が 出来 そう だ けど 、 大丈夫 ?]

correct:[b7 と c4 なら どっち ?]
predicted:[b7 と g2 なら どちら ?]

correct:[どこ が 最善 手 だろ う か]
predicted:[2 番 め に 良い 手 は ?]

Fig. 4　言語生成部のモデルの出力結果

のではないかと考えられる．

4.3　言語生成部の実験について

ユーザーの質問に対する回答のデータを収集することが困難であったため，今回の実験では，言語理解部で使用したデータセットを使用し，質問タイプ，質問キーワードクラス，質問キーワードを入力として，それらに対応するユーザーの質問文が再現出来るかを調べた．生成された文の評価は行えていないので，実際に生成された文を実験結果として提示する．

Fig. 4 に言語生成部の出力例を示す．correct が正しい文で，predicted が出力された文である．2 つ目までの結果は，正しい文に近いものが生成された結果である．テストデータ 183 文中，104 文でこのような結果になった．3 つ目の結果は，文構造としては正しいものが出力されているが,「座標」（スロットの値）が間違っている．テストデータ 183 文中，30 文がこのような結果になった．このような出力が成された原因としては，文構造だけでなく，スロットの値も含めて一度に出力しようとしたことが考えられる．このような間違いは特に「座標」がキーワードとして含まれる文章で多く見られた.「座標」に分類される質問キーワードは a1〜h8 の 64 個あるため，正しく推定出来なかったのではないかと考えられる．4 つ目の結果は，正しい文とは全く違う文が出力された結果である．テストデータ 183 文中，49 文でこのような結果になった．

5　インタビュー発話の生成に向けて

5.1　インタビュー対話データについて

現在は，これまでの研究で構築してきたモデルを，高齢者のインタビュー対話データに適用している．インタビュー対話データは，実際に何らかの社会参加活動に参加している高齢者にインタビューを行っているもので，高齢者（複数人の場合もある）とインタビュアーの発話で構成されている．高齢者の発話は，既にラベル付されており，このラベルを発話意図のフレームとスロットとして取り扱う．このラベルは，高齢者の発話が本人の社会参加に関連しているかどうか，という観点から付けられている．発話意図のフレームに当たるラベルを以下に示す．

- OBJ1: 高齢者が所属しているグループに対する参加に明示的に関連している発話
- OBJ2: 高齢者が所属しているグループに対する参加に明示的に関連していない発話
- OBJ3: 高齢者が所属しているグループの他の社会参加活動に関する発話
- OBJ4: 高齢者が勤めている，もしくは勤めていた会社に関する発話
- OBJ5: 上記のカテゴリに入らない，活動以外のことに関する発話

発話意図のスロットに当たるラベルの例を以下に示す．ラベルの数は 48 である．

- 利己的志向: 自己の楽しみ，生きがい，健康志向などについての内容
- 性格の固さ: 現役時代の地位へのこだわりやプライドについての内容
- ボランティア認知: ボランティアの有用性に関するイメージについての内容
- gender の問題: 男性が女性中心のグループに入れない，など

インタビュアーの発話意図のフレーム，スロットは現在ラベル付を行っている所であり，フレーム，スロットの種類も調整している段階である．インタビュアーの発話意図のフレームに当たるラベルの例を以下に示す．ラベルの数は 8 である．

- 相槌: 相手発話に対して，相槌を入れる．共感や感想なども含まれる．
- 質問: 相手に質問をする．
- 意見: 相手発話に対して，自分の知識や意見，考察を述べる．

インタビュアーの発話のスロットに当たるラベルの例を以下に示す．ラベルの数は 27 である．

- 共感: 相手発話に共感，同意をする．
- 理由: 相手に，出来事や行動に関する理由や原因を質問する．
- 社会参加活動: ボランディア，シニアクラブなど，社会参加活動について言及している．

5.2　言語理解部

高齢者の発話意図のスロットに当たるラベルは，単語や表現単位ではなく，文章単位で付けられているので，3.2 の $Model_{SLU}$ を直接適用出来ない．そこで，高齢者発話を句点，読点で区切り，これらを「一文」として，一文ごとにラベルを対応させる．具体的に

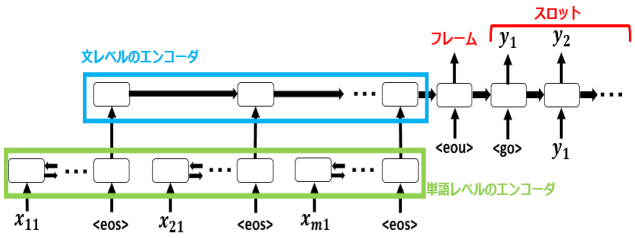

Fig. 5　インタビュアーエージェントの言語理解部モデルの概略図

は，高齢者発話に関して，句点と読点を <eos> に変換し，一文ごとに単語レベルのエンコードを行う．これらのエンコード結果をさらに時系列的にエンコードすることで，高齢者発話全体をエンコードする．このエンコードを実現するために，3.2 の $Model_{SLU}$ のエンコーダ部を階層的にした階層的 Encoder-Decoder を用いる [4].

現在実装しているモデルを Fig.5 に示す．$x_{11}, \ldots, <eos>$ は，<eos> で区切られた，高齢者発話の一文である．これらがまず，単語レベルでエンコードされ，<eos> がエンコードされた時点での LSTM の隠れ状態ベクトルが，文レベルのエンコーダの入力になる．次のステップで，高齢者発話の次の一文 $(x_{21}, \ldots, <eos>)$ がエンコードされ，その結果と，前の文レベルのエンコーダの隠れ状態が現時点での文レベルのエンコーダの入力になる．このようにして，単語レベルと文レベルで階層的にエンコードを行う．高齢者発話の終端に，発話終了を表す <eou> を付ける．高齢者発話のエンコードが終了した次のステップからデコードが始まり，<eou> が入力され，発話意図のフレームが出力される．その次のステップから，時系列的に発話意図のスロットが出力される．

5.3　対話管理部と言語生成部

対話管理部と言語生成部に関しては現在構想段階である．インタビュー対話データにおける，インタビュアー発話のラベル付けがある程度進んだ段階で実装に移る予定である．

対話管理部では，5.1 で定義した高齢者の発話意図からインタビュアーの発話意図への変換を行う．最終的には，回答に必要な情報を得るためのデータベース検索や，インタビュアーのスキル，対話フローの反映などを構想しているが，本研究では，その土台として，ディープニューラルネットワークを用いて，どれだけインタビュアーの発話意図が正しく変換されるかを調査する．ディープニューラルネットワークでの実装，実験の完了後，次のステップとして，対話管理部を Deep Q-learning で構築する [5]．言語理解部では，インタビュアーの発話意図を基に，2.3 で述べた $Model_{SLG}$ を用いてシステムの発話生成を行うことを想定している．

6　おわりに

本稿では、インタビュワーエージェント構築に向けて，これまでの研究である，「オセロゲーム中にユーザーを支援する質問応答システム」において，言語理解部と言語生成部を LSTM Encoder-Decoder で構築し，実験を行った．結果として，言語解析部の実験結果としては，質問タイプの推定率が 85.3%，質問キーワードクラスの推定率が 83.6%という結果になった．結論として，質問タイプ（フレーム）推定を行う場合は単独のモデルを，質問キーワードクラス（スロット）推定を行う場合は $Model_{SLU}$ を用いた方が良い，ということが分かった．これは，*joint* 学習をする時の各モデルの信頼度に依存すると考えられる．また，言語生成部の結果としては，キーワードも含めて言語生成を試みたが，キーワードの部分だけ間違って生成される結果が見受けられたので，キーワードの部分をブランクの状態で言語生成を行い，後でキーワードを埋め込む，という形をとることを検討している．

謝辞　本研究の一部は、JSPS 科研費 JP17K00236, JP17H01995 の助成を受けたものである。

参考文献

[1] 松好祐紀, 滝口哲也, 有木康雄. ”Attention-based LSTM を用いた意図理解とキーワード抽出の統合による質問応答システム ,” 電子情報通信学会技術研究報告, Vol. 118, No. 198, pp. 9-14, 2018-08.

[2] B. Liu and I. Lane, ”Attention-Based Recurrent Neural Net-work Models for Joint Intent Detection and Slot Filling,” in INTERSPEECH. 2016, pp. 685-689

[3] Minh-Thang Luong *et al.* ”Effective Approaches to Attention-based Neural Machine Translation, ” arXiv preprint arXiv:1508.04025, 2015

[4] Ryo Masumura *et al.* ”Online Call Scene Segmentation of Contact Center Dialogues based on Role Aware Hierarchical LSTM-RNNs, ” Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference 2018

[5] Pei-Hao Su *et al.* ”Continuously Learning Neural Dialogue Management, ” arXiv preprint arXiv:1606.02689v1, 2016