

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¹, 滝口哲也¹, 有木康雄¹, 瀧和男², 片桐恵子³

¹神戸大学 システム情報学研究科, ²神戸大学 科学技術イノベーション研究科,

³神戸大学 人間発達環境学研究科

yuki.matsuyoshi@stu.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp,
taki@port.kobe-u.ac.jp, katagiri_k@people.kobe-u.ac.jp

1 はじめに

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。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。

本研究では、言語理解部、言語生成部に関して、ベースとなるシステム「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]を用いている。このシステムは、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るものである。この質問応答システムにおいても、インタビュー対話システムにおいても、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、ユーザーに対する発話生成を行うことは共通している。従って、まず [1] の研究で構築したシステムについて述べ、次に、このシステムをインタビュー対話に用いる方法について、5. 以降で述べる。

2 これまでの研究

システムの構成としては、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。言語理解部ではユーザーの発話意図を推定し、これを対話管理部に送る。対話管理部では、ユーザーの発話

意図を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。このシステムにおける言語理解部、言語生成部について述べる。

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。フレームはユーザーのたまかな意図を表す「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」ならば、質問キーワードクラスは「オセロ用語」になる。質問タイプは、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
- 勝敗: 勝敗や形勢に関する質問

質問キーワードクラスは、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 言語理解部

言語理解部は図 1 のような、LSTM Encoder-Decoder で実装している（以下、 $Model_{SLU}$ とする）。ユーザーからの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 $(x_1 \dots x_m)$ に変換して word embedding を行い、LSTM に入力する。文末記号 $\langle eos \rangle$ 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が出力され、 $\langle go \rangle$ が入力された後、

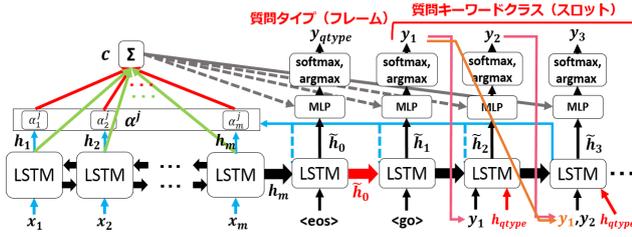


図 1: $Model_{SLU}$ の概略図

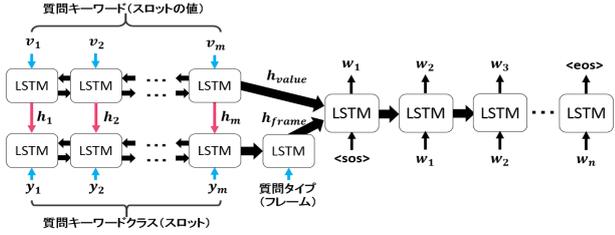


図 2: $Model_{SLG}$ の概略図

質問キーワードクラス (y_1, y_2, \dots, y_j) が順番に出力される [2]. また, このモデルでは Attention 機構を導入している [3]. 各デコードステップの隠れ層ベクトル $\tilde{h}_j (j=0, 1, 2, \dots, m)$ と, 質問文入力時に保持しておいた, エンコーダの隠れ層ベクトル $h_i (i=1, 2, \dots, m)$ を利用して, 図 2 の α_i^j を計算する. j はデコードステップを表す.

$$\alpha_i^j = \frac{\exp(\mathbf{W}_1^T \tanh(\mathbf{W}_2 \mathbf{h}_i + \mathbf{W}_3 \tilde{\mathbf{h}}_j))}{\sum_{n=1}^m \exp(\mathbf{W}_1^T \tanh(\mathbf{W}_2 \mathbf{h}_n + \mathbf{W}_3 \tilde{\mathbf{h}}_j))} \quad (1)$$

$\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$ は学習するパラメータである. α_i^j を h_i の重みとし, コンテキストベクトル \mathbf{c} を計算する.

$$\mathbf{c} = \sum_{i=1}^m \alpha_i^j \tilde{\mathbf{h}}_j \quad (2)$$

$\mathbf{c}, \tilde{\mathbf{h}}_j$ から, 以下の式により y_{qtype}, y_j を計算する. $\mathbf{W}_4, \mathbf{W}_5$ は学習するパラメータである.

$$y_{qtype} = \operatorname{argmax}(\operatorname{softmax}(\tanh(\mathbf{W}_4 \mathbf{c} + \mathbf{W}_5 \tilde{\mathbf{h}}_0)))$$

$$y_j = \operatorname{argmax}(\operatorname{softmax}(\tanh(\mathbf{W}_4 \mathbf{c} + \mathbf{W}_5 \tilde{\mathbf{h}}_j))) \quad (3)$$

$Model_{SLU}$ の特徴は, デコーダでの各時間ステップにおける入力に 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 $j-2$ の出力 y_{j-2} も利用している.

3.3 言語生成部

言語生成部では, システムの発話意図からシステムの発話を生成する. モデルは図 2 のような Encoder-Decoder で構築している ($Model_{SLG}$). エンコーダは,

表 1: 質問タイプ, 質問キーワードクラス推定に関する結果. (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

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, $\langle \text{sos} \rangle$ 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 $(h_1 \dots h_m)$ 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 $\mathbf{W}_6, \mathbf{W}_7$ を用いて, h_{value} と h_{qtype} を連結したベクトルである.

$$\mathbf{W}_6 h_{value} \oplus \mathbf{W}_7 h_{qtype} \quad (4)$$

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. 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.

4.2 言語理解部の実験結果

実験結果を表 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のではないかと考えられる.

```

correct:[切り返して どういう こと ですか ?]
predicted:[切り返して 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 ? ]

```

図 3: 言語生成部のモデルの出力結果

4.3 言語生成部の実験について

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、ユーザーの質問文が再現出来るかを調べた。

図3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。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。このような出力が生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LSTM Encoder-Decoderで一度に出力しようとしたことが考えられる。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、104文で正しい文に近いものが生成された。

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

質問応答システム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する方法について述べる。インタビュー対話データは、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の他の社会参加活動に関する発話

- OBJ3: 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
- OBJ4: 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、スロットのラベル付けを現在行っている。このラベル付けは、複数人で人手で行っている。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: ボランティア、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文レベルのエンコードを行うことで、高齢者の1発話をエンコード

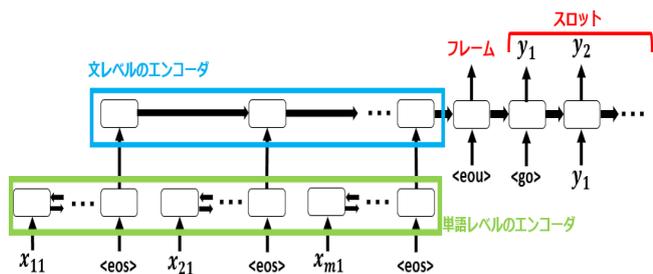


図 4: 高齢者発話の言語理解部モデルの概略図

する。このエンコードの実現に、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図 5 に示す。 $x_{11}, \dots, < eos >$ 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 ($x_{21}, \dots, < eos >$) 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、現在、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している [5]。言語生成部では、インタビュアーの発話意図を基に、3.3 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 終わりに

本稿では、インタビュアーエージェント構築に向けて、ベースとなる研究である「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。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% という結果が得られた。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。また、言語生成部の結果としては、キーワードの部分だけ間違っ生成される結果が見受けられたので、キーワードの部分ブランクの状態と言語生成を行い、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。

インタビュー対話については、インタビュアーの発話に対してラベル付けを完了させ、言語生成の実験と対話管理の実装を行う予定である。

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 Network 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