

ニューラル質問応答モデルの仮想世界から現実世界のデータへのドメイン適応

宮西大樹^{*1} 川鍋一晃^{*2}

^{*1*2}国際電気通信基礎技術研究所

^{*2}理化学研究所 革新知能統合研究センター

1. はじめに

近年の行動認識術の発展により、日常生活行動を表すセンサデータを時系列に並ぶ言語データに変換できるようになった [Krishna 17, Miyanishi 18]. このような日常生活を表す言語データを読み込んで、実世界の内容を問う様々な質問に解答できる実世界質問応答システムが実現すれば、現実世界の内容を理解するロボットや、室内での物探し、患者の医療モニタリングや個人の記憶補助に役立てることができる。

実世界質問応答の実現には、ユーザーによって与えられる質問と現実世界を表す時系列に並んだ複数の文（ストーリー）の関係から推論を行い、解答文となる単語または単語列を生成する必要がある。例えば、「サイトーさんがキッチンでコーヒーを作り、コーヒーを手を持った後、リビングに移動して、コーヒーを机の上に置いた。そしてサイトーさんは TV を付けた後にコーヒーを飲んだ。」というストーリーと「いまサイトーさんはどこにいる？」という質問文が与えられたとき、質問応答システムは「リビングルーム」という単語を出力する。本研究では、外部メモリの機構を持つニューラルネットワーク（以下、ニューラル質問応答モデル） [Kumar 16, Xiong 16] を用いてこの質問応答の課題に取り組む。

ニューラル質問応答モデルは、高精度な質問応答が可能であるものの、モデルの性能を発揮するために、質問・ストーリー・解答の三つ組の学習データが大量に必要である。しかし、実世界のデータはプライバシーの問題やラベリングに多大な労力を必要とするため、大量の学習データを用意することは困難である。

そこで、本研究ではライフシミュレーションで取得した仮想世界のデータを用いて質問応答の学習データを作成し、このデータで学習したニューラル質問応答モデルを敵対的損失関数を用いて現実世界のデータにドメイン適応させることで、この問題に対処する。我々はライフシミュレーターとして The Sims 4 (図 1 参照) を使い仮想世界の質問応答データを取得し、住宅環境内で収集した日常生活行動に関する現実世界の質問応答データで評価を行う。

2. 実世界質問応答

本節では、実世界の内容を質問して解答を生成するニューラル質問応答モデルについて紹介する。図 2 に本手法の枠組みを図示する。提案手法では、質問と文書のエンコーダを仮想世界のデータ（ソース）と現実世界のデータ（ターゲット）の 2 種類に分け、仮想世界のデータで学習したエンコーダと解答を出力するデコーダを用いてソース側のデータで学習を行い、次にソース側のデータとエンコーダ、ターゲット側のデータとエンコーダを



図 1: ライフシミュレーションゲーム The Sims 4 のプレイ画面。ゲーム内のキャラクターは自身の性格や欲求に従って、室内を半自動的に行動する。

用いて、敵対的損失を通してドメイン適応を行う。最後に、ドメイン適応したターゲット側のエンコーダをソース側のデコーダと組み合わせて質問応答システムを構成して、評価を行う。本研究では、エンコーダとして動的記憶ネットワーク (Dynamic Memory Networks; DMN)、デコーダとしてゲート付き回帰ユニット (gated recurrent unit; GRU) を用いる。質問応答システムは、質問 Q とストーリー $\mathbf{D} = [D_1, D_2, \dots, D_N]$ が与えられたとき、解答 A を返す。ここで、 $Q = (w_1^Q, w_2^Q, \dots, w_n^Q)$ を質問文の単語、 $\mathbf{D} = (w_1^D, w_2^D, \dots, w_n^D)$ をストーリー中の文の単語、 $A = (w_0^A, w_1^A, \dots, w_n^A)$ を解答文中の単語に対応する単語ベクトルとする。

2.1 デコーダ：ゲート付き回帰ユニット

エンコーダの出力から解答となる単語列を生成するため、下式で定義されるゲート付き回帰ユニット [Chung 14, Cho 14] (以下、GRU と略記) を用いる。

$$\begin{aligned} z_i &= \sigma(W_z x_i + U_z h_{i-1} + b_z) \\ r_i &= \sigma(W_r x_i + U_r h_{i-1} + b_r) \\ \tilde{h}_i &= \tanh(W x_i + r_i \circ U h_{i-1} + b_h) \\ h_i &= (1 - z_i) \circ h_{i-1} + z_i \circ \tilde{h}_i \end{aligned}$$

ここで、 x_t は時刻 t の入力ベクトル、 h_t は状態ベクトル、 σ はシグモイド関数、 $\circ$ はアダマール積を表す。また、 $W_z, W_r, W \in \mathbb{R}^{n_H \times n_I}$ と $U_z, U_r, U \in \mathbb{R}^{n_H \times n_H}$ 、次元 n_I は入力ベクトルの大きさ、次元 n_H は状態ベクトルの大きさ、 b_z, b_r, b_h はバイアス項を表す。以下、上記 GRU を表す関数を $h_t = \text{GRU}(x_t, h_{t-1})$ と略記する。解答文を生成する際は、単語列の生成確率 $P(A) = \prod_{t=1}^{T+1} P(w_t^A | A_{[0, t-1]})$ に基づいて単語のサンプリングを行う。ここで、 $P(A_t | A_{[0, t-1]}) = \text{softmax}(W_o h_t^A + b_o) \cdot w_t^A$ 、 $h_t^A = \text{GRU}(w_t^A, h_{t-1}^A)$ 、 $W_o \in \mathbb{R}^{n_V \times n_H}$ 、次元 n_V は語彙数となる。

連絡先: 宮西大樹, 国際電気通信基礎技術研究所, 京都府相楽郡精華町光台 2 丁目 2-2, miyanishi@atr.jp

2.2 エンコーダ：動的記憶ネットワーク

質問とストーリーをベクトル表現に変換するため、エンコーダとして、動的記憶ネットワーク（以下、DMN と略記）[Xiong 16] を用いる。DMN は、入力部、エピソード記憶部、質問、解答部の4つに分かれる。本稿では、解答部は前述のデコーダを用い、入力部は、ストーリー中の文のベクトル表現を事実情報 f_i とし、複数の事実間の関係を GRU を用いて保持し、質問文のベクトル表現と注意機構 [Bahdanau 15, Luong 15] を用いて質問に対応する内容を問い合わせる。

2.2.1 入力部と質問

DMN の入力部では、ストーリーの各文をベクトル表現に変換し、それらの前後関係を表現する。文のベクトル表現を得るため、文中の単語の位置を考慮した位置埋め込みを行う。具体的には、 $s_i = \sum_{j=1}^J l_j \circ w_j^i$ 、ここで $\circ$ は要素ごとの積を表す。 l_j は列ベクトルであり、 $l_{jd} = (1 - j/J) - (d/n_v)(1 - 2j/J)$ とする。 d は単語の文中での位置、 J は文中の単語の数を表す。質問文のベクトル表現は $q = \sum_{j=1}^J l_j \circ w_j^Q$ 、ストーリー中の文のベクトル表現は、 $d = \sum_{j=1}^J l_j \circ w_j^D$ になる。

次に、ストーリー中の文間の双方向の関係を捉えるため、双方向 GRU (Bidirectional GRU; bi-GRU) を用いる。

$$f_i = \text{bi-GRU}(f_{i-1}, f_{i+1}, d_i) \quad (1)$$

ここで、 f_i は i 番目の双方向 GRU の状態ベクトル、 f_{i-1} は前向き GRU の状態ベクトル、 f_{i+1} は後ろ向き GRU の状態ベクトルになる。状態ベクトル f を事実とし、これがエピソード記憶部の入力となる。

2.2.2 エピソード記憶部

DMN のエピソード記憶部では、質問に応じて特定の事実に注目しながら $F = [f_1, \dots, f_N]$ の検索を行う。ここで、注意ゲート g_i^t は、質問 q が事実 f_i に対してどの程度注意するかを表す。

$$z_i^t = [f_i \circ q; f_i \circ m^{t-1}; |f_i - q|; |f_i - m^{t-1}|] \quad (2)$$

$$Z_i^t = W_\beta \tanh(W_\alpha z_i^t + b_\alpha) + b_\beta \quad (3)$$

$$g_i^t = \frac{\exp(Z_i^t)}{\sum_{k=1}^{J_i} \exp(Z_k^t)} \quad (4)$$

ここで、 f_i は i 番目の事実で、 m^{t-1} は以前のエピソード記憶、 q は質問文のベクトル表現、 $|\cdot|$ は要素ごとの絶対値、 $;$ はベクトルの結合を表す。

次に注意ゲート g_i^t と GRU を用いて、質問 q と事実 F の関係を表す文脈ベクトル c を求める。ここで、GRU の式 (1) の z_t を g_i に置き換えることで、質問 q に応じて事実 f_i の前後の文脈を重み付けした表現を獲得する。

$$h_i = (1 - g_i^t) \circ h_{i-1} + g_i^t \circ \tilde{h}_i \quad (5)$$

GRU の h_N を文脈ベクトル c とする。

記憶の更新は t 回行う。 t 回目文脈のベクトル表現を c^t 、 t 回更新後のエピソード記憶を表すベクトル表現を m^t とする。記憶の更新は下式で行う。

$$m^t = \text{ReLU}(W^t[m^{t-1}; c^t; q] + b) \quad (6)$$

ここで、 $W^t \in \mathcal{R}^{n_H \times n_H}$ 、活性化関数として ReLU (Rectified Linear Unit) 関数 [Nair 10] を用いる。 m^t がエンコーダ DMN の最終的な出力になる。

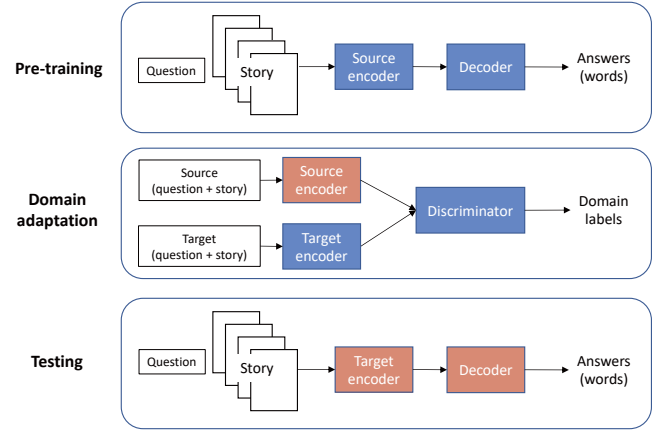


図 2: 敵対的ドメイン適応の模式図

2.3 ニューラル質問応答モデルのドメイン適応

ニューラル質問応答モデルの仮想世界のデータから現実世界のデータへのドメイン適応を行うため、敵対的ドメイン適応 (Adversarial Discriminative Domain Adaptation; ADDA) [Tzeng 17] を用いる。ドメイン適応のため、ADDA では、下記目的関数を段階的に最適化する。

$$\begin{aligned} \min_{M_s, C} \mathcal{L}_{seq}(\mathbf{X}_s, Y_s) &= -\mathbb{E}_{(\mathbf{x}_s, y_s) \sim (\mathbf{X}_s, Y_s)} \sum_{k=1}^K \mathbf{1}_{[k=y_s]} \log C(M_s(\mathbf{x}_s)) \\ \min_D \mathcal{L}_{adv_D}(\mathbf{X}_s, \mathbf{X}_t, E_s, E_t) &= -\mathbb{E}_{\mathbf{x}_s \sim \mathbf{X}_s} [\log D(M_s(\mathbf{x}_s))] - \mathbb{E}_{\mathbf{x}_t \sim \mathbf{X}_t} [\log(1 - D(M_t(\mathbf{x}_t)))] \\ \min_{M_s, M_t} \mathcal{L}_{adv_M}(\mathbf{X}_s, \mathbf{X}_t, D) &= -\mathbb{E}_{\mathbf{x}_s \sim \mathbf{X}_s} [\log D(M_t(\mathbf{x}_t))] \end{aligned} \quad (7)$$

ここで、 $\mathbf{x}_s$ はソース側のストーリーと質問 $\mathbf{D}_s, Q_s, y_s$ は解答 A_s に相当する。 $\mathbf{x}_t$ はターゲット側のデータを表す。まず、ソース側のエンコーダ M_s とデコーダ C をラベルありのデータ $\mathbf{X}_s$ と Y_s を用いて $\mathcal{L}_{seq}$ を最適化する。次に、ソース側のエンコーダ M_s の重みを固定して、ターゲット側のエンコーダと識別器 D を用いて $\mathcal{L}_{adv_D}$ を最適化する。その後、デコーダ M_t の $\mathcal{L}_{adv_M}$ を最適化する。図 2 にドメイン適応の模式図を示す。

3. データセット

3.1 仮想世界のデータ：The Sims のシミュレーション

日常生活のシミュレーションデータを取得するため、シミュレーションゲーム The Sims 4 を用いる。The Sims 4 では Sim と呼ばれる人間を模したキャラクターが自身の性格や感情や欲求に従って屋内を半自動的に行動する。例えば、空腹ゲージが下がるとキッチンの冷蔵庫から食べ物を取り出して、ダイニングの机に置いて食事を始めたり、膀胱ゲージが下がるとトイレに行く行動を取りやすくなる。Sim が生活する屋内環境はカスタマイズ可能であり、物体（コーヒーメーカー、食洗機、PC など）を追加することで Sims の行動が変化する。また、物体を選択することで、その物体に関連する行動を Sim にとらせることも可能である。

本研究では、日常生活行動のシミュレーションを行うため、異なる 3 つの住宅環境を用意した。各環境には、キッチン・ダ



図 3: The Sims 4 の屋内環境の例, 左図は男性とペットの住環境, 右図は夫婦の住環境.

イニング, リビング, バスルーム, ベッドルームの部屋と, 日常生活に必要な物体 (冷蔵庫やベッドや TV など) を用意した. 各環境内では, 男性 1 人とペット 1 匹 (犬), 男性 2 人, 男性 1 人女性 1 人の夫婦とその子供一人 (女性) の 3 つの組み合わせそれぞれ日常生活を送る. 日常生活のシミュレーションでは, 1 日の行動を 1 セッションとして, 10 セッション分のデータを取得し, 各個人にどこで・だれが・何をしたかといった日常生活行動を表す文のラベリングを行った.

3.2 現実世界のデータ: 屋内での日常生活行動

提案手法の評価を行うため, 日常生活行動データセットを用いる. このデータセットは日常生活行動をヘッドマウントカメラで撮影し, 撮影した一人称映像をもとにして, いつ・どこで・誰が・何をしたかを表す文を手でラベリングしたものである. このデータセットを作成するため, 5 人の被験者がモーションセンサとウェアラブルカメラを装着して, 19 個の日常生活行動を異なる場所ごと固有の行動をした. 被験者はワークシートに書かれた行動を被験者の意思に従って任意のタイミングで行う. 例えば, 被験者はソファに座ってリビングルームでテレビを見ながらコーヒーを飲み, その後, 台所に移動して皿を洗う, といった連続した日常生活行動を行う. この一連の行動を 1 セッションとして 10 回繰り返し, 1 人につき 10 セッション分のデータを収集した. このセミナチュラルな実験プロトコルによるデータ収集 [Bao 04] は, 限られた環境の中で多様な行動についてのセンサデータを集めるために, 行動認識の研究でよく用いられる. 質問応答のデータ・セットを用意するため, 質問のテンプレートを用意し, 仮想世界と現実世界の文に対応する解答をルールベースで作成した.

4. おわりに

本稿では, ニューラル質問応答モデルのドメイン適応手法と当該手法の有効性を検証するためのデータセットを紹介した. 実験設定の詳細や実験結果については, インタラクティブ発表にて紹介する.

5. 謝辞

本研究は, JSPS 科研費 16K21718, JST CREST JP-MJCR15E2 の助成を受けたものです.

参考文献

[Bahdanau 15] Bahdanau, D., Cho, K., and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, *ICLR* (2015)

[Bao 04] Bao, L. and Intille, S. S.: Activity recognition from user-annotated acceleration data, in *Pervasive*, pp. 1–17 (2004)

[Cho 14] Cho, K., Merriënboer, van B., Bahdanau, D., and Bengio, Y.: On the Properties of Neural Machine Translation: Encoder–Decoder Approaches, in *SSST-8* (2014)

[Chung 14] Chung, J., Gülçehre, Ç., Cho, K., and Bengio, Y.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, in *Deep Learning and Representation Learning Workshop* (2014)

[Krishna 17] Krishna, R., Hata, K., Ren, F., Fei-Fei, L., and Niebles, J. C.: Dense-Captioning Events in Videos, in *ICCV*, pp. 706–715 (2017)

[Kumar 16] Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., and Socher, R.: Ask me anything: Dynamic memory networks for natural language processing, in *ICML* (2016)

[Luong 15] Luong, M.-T., Pham, H., and Manning, C. D.: Effective approaches to attention-based neural machine translation, *EMNLP* (2015)

[Miyanishi 18] Miyanishi, T., Hirayama, J.-i., Maekawa, T., and Kawanabe, M.: Generating an Event Timeline about Daily Activities from a Semantic Concept Stream, in *AAAI* (2018)

[Nair 10] Nair, V. and Hinton, G. E.: Rectified linear units improve restricted boltzmann machines, in *ICML*, pp. 807–814 (2010)

[Tzeng 17] Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T.: Adversarial discriminative domain adaptation, in *CVPR* (2017)

[Xiong 16] Xiong, C., Merity, S., and Socher, R.: Dynamic memory networks for visual and textual question answering, in *ICML*, pp. 2397–2406 (2016)