Ontology DrivenGAN と対話アニメーションの構想

長内洋太[†1]　森山紘行[†1]　李亜超[†1]
下川原（佐藤）英理[†1]　山口亨[†1]

概要：近年，人とロボットの間のコミュニケーションを円滑に進めるために，様々なアプローチが検討されている．それらのロボットは，言語や映像を意思伝達の手段として用いるものから，ジェスチャーのようなロボットの物理的動作を用いるものまで様々である．しかし，人とロボットのコミュニケーションにおいて，単一の入出力に頼ることは，それ以外の手段で汲み取れるはずの情報を切り捨てている危険があり，それは，ロボットが求められる応答をできない可能性がある．本研究では，コミュニケーションにおける 1 場面のオントロジーをまとめて入出力に用いて並列に処理を行うことで，人がより良いレスポンスを得ることを目的としている．本論文では，入力が複数であった場合において，どのような出力が理想的であるかの予備実験・考察を行い，並列処理の可能性について検討したものである．

1. はじめに

近年，人とコミュニケーションを行うロボットは，技術の発達によって目覚ましい発展を遂げている．それらは言語情報のみならず，画像や映像を用いたり，ジェスチャーを読み取ったり，心拍などの生体情報を用いるものまで様々である．ここで，ロボットが汲み取る人間側の情報を指定した場合，本来人間同士がコミュニケーションを行うときと異なり，指定した情報のみではロボットが人間側の情報を取りこぼし，期待されたレスポンスを行えないという危険性を孕む問題が存在する．

本研究では，人間とロボットのインタラクションにおける 1 場面の様々な情報を入力としてシステムを構築し，並列処理することを試みる．今回はその予備実験として，様々な入力の形式について，敵対的ニューラルネットワークを用いて学習させ，それぞれの出力結果を考察した．

2. 提案手法

人間とインタラクションを行うロボットのモデルの一部に GAN を採用し，人間の振る舞い[1]や，対話行為タグによる対話の制御[2]について検証を行った研究はいくつか存在するが，対話行為における複数の情報について，並列に処理を試みた例は存在しない．

情報科学において，オントロジーという考え方が存在する．オントロジーは，個体（インスタンス）や概念（クラス），属性などの様々な情報と，その関係を記述する概念体系である．本研究では，人間とロボット間の対話における 1 場面を，オントロジーの概念に落とし込む．オントロジーを入力として用いる場合，入力の形式は複数かつ多様な形式となる．また，利用者がどのような場面でロボット（システム側）とコミュニケーションを試みるか，ロボットにどのような返答を期待しているかという前提によって，想定される入力は様々である

今回は，複数の GAN を用いて複数の異なる形式のデータを並列処理し，それぞれから出力を得て統合，ユーザとの対話インタラクションに利用する Ontorogy DrivenGAN（図 1）についての構想とその実現可能性，課題について検証する．

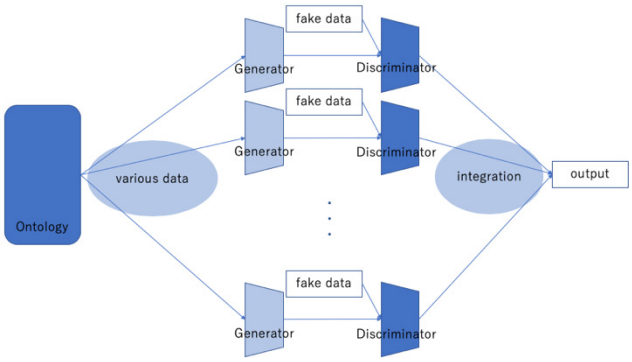

図 1. Ontorogy DrivenGAN のシステム全体の構成

なお本実験では，GAN における様々な出力の検証と我々の先行研究に照らし合わせ，入力するデータ（図 1 の“various data”）を“画像（picture）”，“人の動作（character）”，“文（text）”の 3 つに絞り，図 2 のような構成を想定した．この構成図において，扱うデータごとに個別に GAN を用いて入出力を行う予備実験を行い，その結果と考察について述べる．

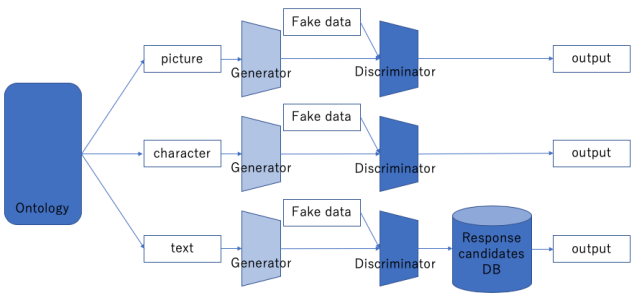

図 2. 今回の実験のシステム全体の構成

†1 首都大学東京大学院

© 2020 Information Processing Society of Japan

3. 予備実験

3.1 対話応答候補文の生成

ロボットが人間と言語を用いて実対話を行い，コミュニケーションを行う研究は，近年盛んに行われている．対話システムは主に，道案内や商品説明，Web 検索等のユーザが特定の目的を持って開始するタスク指向型（task-oriented）[3]と，チャットボット等，ユーザが特定の目的を持たず雑談のために行う非タスク指向型（non task-oriented）[4]の 2 つに分けることができる．

非タスク指向型対話システムにおいて，応答の候補となる文を大量にデータベース内に保持し，文選択アルゴリズムを用いて，ユーザ発話に適した応答文を出力するシステムが研究されている[5]．このシステムは，応答文を出力する際にスコアリングやフィルタリングを行うことにより，非文を出力しにくいという利点があるが，データベース内の文量が少ないときに出力の精度の問題が発生する．我々はこの問題について以前から研究を行ってきたが[6][7]，応答文の生成をルールベースな手法によって行うと，ユーザの，ロボットとの対話への興味について，問題が生じることが分かっている．

今回は，先行研究で行ったルールベースな手法と対比を行うため，敵対的ニューラルネットワークである GAN[8]にモンテカルロ探索を適用し，勾配伝搬を行うことで文生成に用いることが出来るようにした，Lantao Yu らが提唱した seqGAN[9]を用いて日本語対話応答文の生成を行い，その出力と，対話応答文の生成に seqGAN を用いる妥当性についてそれぞれ考察した．

3.1.1 実験

今回は，ロボットとユーザ間の雑談対話における，ユーザ発話 1,501 文を書き起こした in-house なデータを用意した．前処理として入力データを形態素解析器 Mecab[10][11]を用いて文を品詞に分割し，その後学習と文生成，出力の検証を行なった．Discriminator の学習率は 1e-4，Generator の最適化には Adam を用いた．

3.1.2 実験結果と考察

上記の条件で実験を行った結果，得られた文の例を表 1 に示す．

本実験で生成した文について，例から分かるように，主語が入っていないものがほとんどであることが分かった．これは，Generator の pre-train に用いたユーザ発話文が，ロボットとの 1 対 1 対話における発話であり，自らの状況について述べている文が多かったたことが影響していると考えられる．また，表 1-(1)のように文法として誤っている文や，表 1-(2)のように文法的に正しくとも，動詞がその名詞に適していない文が散見された．これらの問題を解決するために，出力文が文法的に正しいか否

表 1. epoch 数と生成文

epoch	生成文の例
1	ある程度自然素材のものを買ってみました。
	今年を挑戦するのが行きますね。(1)
10	献立でコーヒーは毎日食べています。(2)
	クラッシクギターを捨てました。
50	お米もよく食べています。
	町内会のイベントをして掃除をして楽しいですよね。
100	ホームベーカリー、全然で、心配をしています。
	ネットでよく読んで行きました。

かの判定に係り受け解析器等を利用する，pre-train 時のデータ数を増やすことで，対象の名詞と動詞の共起の回数を増加させる等の手法が考えられる．今後はこれらの解決策の導入の検討とともに，生成文の客観的な評価を行なっていきたい．

3.2 アニメーション背景の生成

近年，スタイル変換に関する研究が多く行われてきており，多くの場合 Ian Goodfellow らが提唱した GAN[8]を基本として用いている．この派生系となる研究は数多くあり，例えばある画像から別の画像スタイルへの変換では CartoonGAN[12]と呼ばれる生成手法により，元の画像を漫画風スタイル画像に変換することが可能になっている．また，動画から動画へのスタイル変換タスク[13]なども出てきている．

動画は，連続した画像フレームによって成り立っている。そこで CartoonGAN の Generator を各画像フレームの生成として用いた。動画と画像の違いは，動画は空間特徴だけでなく，時間的特性をも各フレーム間に持っていることである．我々の研究目標は滑らかなアニメーション生成であるため，次の 3.1 節で述べるようなオプティカルフローを時間的特性として用い，図 2 で示すシステム構成によるアニメーション生成を行った．

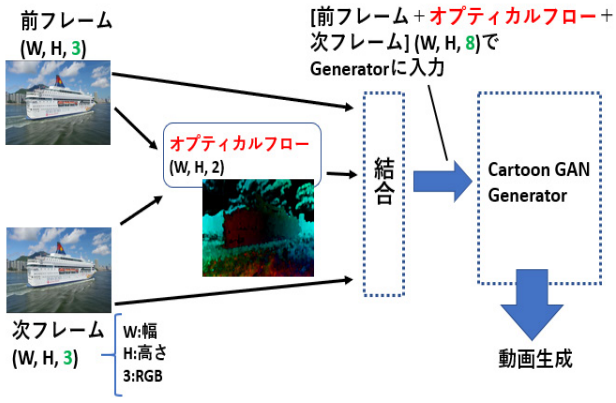

図 3. 本実験でのモデル構成

© 2020 Information Processing Society of Japan

3.2.1 実験

実験に向け，我々は現実世界の動画とアニメーション動画のデータセットを用意した．これらは youtube の中からそれぞれ人手で集めた．今回提案したネットワーク（図 3）と用意したデータセットで実験を行なった．

3.2.2 実験結果と考察

実験を行なった結果，以下の図 3 のようなアニメーションの背景動画を出力することが出来た．本実験で得られた結果について，漫画風にスタイル変換された動画を得られたが，全体として色にムラがあることが確認された．これは，オプティカルフローの弱点を GAN の効果によって消そうとしている可能性が考えられる．そのため，今後は生成された動画の色にムラがある状態を，どうすれば改善出来るか検討し，評価実験も行っていきたい．

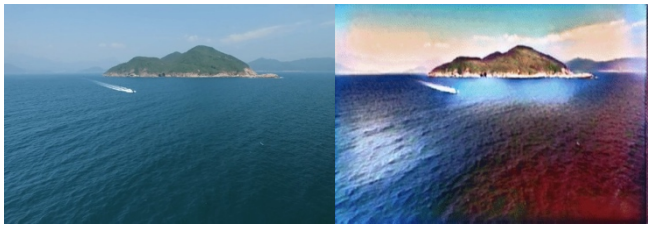

図 4. 元の画像（左）と生成画像（右）

3.3 キャラクターの生成

スタイル変換に関わる研究は近年多く行われてきている．例えば画像においては，pix2pix[14]のように手書きイメージから実際のイメージへのスタイル変換を可能にしているものがある．また，動画のタスクでは，ReCoNet[15]を用いて現実世界の動画から有名な絵画へのスタイル変換を行なっているものがある．

現実世界の動画とアニメーション動画との関係を取得するため，我々は，動きの情報として人の骨格の動きを用いることを提案する．そのためのネットワークを以下の図 5 にて示す．学習プロセスに向けて，まずアニメーションキャラクターの骨格の動きの情報を取得するため，人のポーズを予測するネットワークを用いている．その後，取得した骨格の動きの情報を用いて，アニメーションキャラクターの骨格の動きの生成を行なった．

今回のスタイル変換に伴う推論において，我々は人のポーズを予測するネットワークを現実の人の骨格の動きを取得するのに用い，その動きを GAN における Generator に入力し，アニメーションキャラクターを生成した．

3.3.1 実験

今回は，アニメーションキャラクターのデータセットとして，ゲーム：“NARUTO-ナルト- 疾風伝 ナルティメットストーム 2”のキャラクターの動作を取得したデータセット

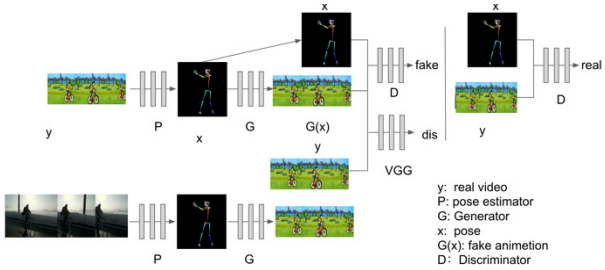

図 5. 本実験でのモデル構成

を用いた．また，背景を取り除きキャラクターを保存しておくため，Mask R-CNN[16]を利用している．この処理によって得られたデータの例を図 6 に示す．

図 6. 元のデータ（左）と取得したキャラクター（右）

3.3.2 結果と考察

本実験では，人間のポーズ推定に Openpese を利用し，Generator は pix2pixHD に基づいて構築した．2 枚の GPU (GTX 1080 TI) を用いた 2 日間の学習によって，円滑な 2 次元のアニメーションキャラクターへの変換を可能とした．出力例を図 7 に示す．

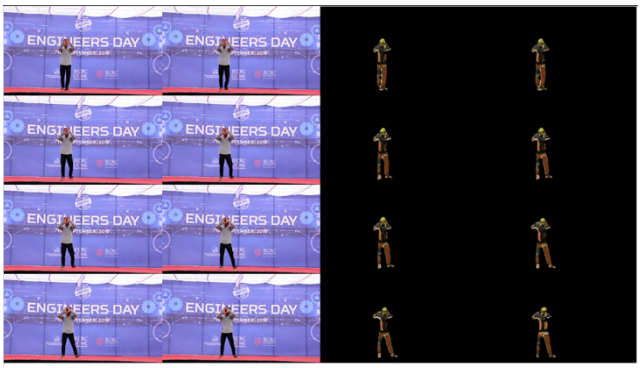

図 7. 元の入力フレーム画像（左）と出力（右）

実験の結果から，

1. 入力の映像から人物の動きを抽出することはできたが，顔や手，足の一部の向きの精度に問題がある．

© 2020 Information Processing Society of Japan

2. 2 次元画像を入力としているため，出力も 2 次元となってしまう．そのため 3 次元の出力には対応できない．

の 2 点が大きな問題として存在することが分かった．一部の部位の精度に関する問題は GAN の機能の強化，例えば FaceGAN[17]を導入することで，正確な顔の出力を得られる可能性がある．また，3 次元のデータを学習データに用いることで，3 次元画像への対応も検討している．

4. まとめと今後の課題

我々は対話アニメーションにおいて，背景，キャラクター，対話文をオントロジーなデータとして 1 つにまとめ，同時に入力，生成する手法を提案した．今回の実験では，個別に 3 種類の GAN を用いて，各データを個別に処理し，人間とロボットの対話行為における 1 オントロジーに含まれる複数の異なるデータを並列して処理する Ontology DrivenGAN の実現に向けて，それぞれの出力についての考察を行った．

我々の構想において，本来入力として想定しているのはオントロジーを考慮した複数形式のデータであり，形式の異なったデータ群について扱う実際のモデルは，1 つのシステムとして統合され，入出力を行う必要がある．
現状の課題としては，

1. どのように各出力を統合し，1 つのオントロジーとして出力するか
2. どのような場面での利用を想定するのか，オントロジーデータをどのように取得するのか
3. その想定した場面において，どのような出力が評価を得られるのか（評価指標）

の 3 つが挙げられる．
今後，これらの問題点についての検討を重ね，システムの構築，出力についての評価を行う予定である．

参考文献

[1] Yusuke Nishimura, Yutaka Nakamura and Hiroshi Ishiguro, Human behavior modeling during interaction using Generative Adversarial Networks, MPS-117, 2018, No.6, pp.1-6.
[2] Seiya Kawano, Koichiro Yoshino and Satoshi Nakamura, An Investigation of Controllable Neural Conversation Model with Dialogue Acts, JSAI Proceedings, 3Rin2-27.
[3] Jason D Williams and Steve Young, Partially Observable Markov Decision Processes for Spoken Dialog Systems, Computer Speech & Language, 2007, vol.21, No.2, pp.393-422.
[4] Richard S Wallace, The Anatomy of ALICE, Parsing the Turing Test, 2009, Springer, pp.181-210.
[5] Ryuichiro Higashinaka, Rashmi Prasad and Marilyn A Walker, Learning to Generate Naturalistic Utterances Using Reviews in Spoken Dialogue Systems, Proceedings of ACL, 2006, pp.265-272.
[6] Youta Osanai, Tomoya Ogata, Mamoru Komachi, Eri-Sato-Shimokawara, Kazuyoshi Wada, Toru Yamaguchi and Yomoya Takatani, Augmentation of Dialogue Database by Generating Interrogative Sentences Using Templates, JSAI Proceedings, 2018, pp.1772-1775.
[7] Youta Osanai, Eri Shimokawara and Toru Yamaguchi, Correlation Analysis of Template Generality and Output Evaluation in Dialogue Response Generation, Proceedings of SCIS&ISIS, 2018, pp.1133-1137.
[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, Generative adversarial nets, NIPS Proceedings, 2014, pp.2672-2680.
[9] Lantao Yu, Weinan Zhang, Jun Wang and Yong Yu, SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient, arXiv, 2017, 1609.05473.
[10] https://taku910.github.io/mecab/
[11] https://github.com/neologd/mecab-ipadic-neologd
[12] Y. Chen, Y.-K. Lai and Y.-J. Liu, CartoonGAN: Generative Adversarial Networks for Photo Cartoonization, Proceedings of Cvpr, 2017, pp9465-9474.
[13] D. Bashkirova, B. Usman and K. Saenko, Unsupervised Video-to-Video Translation, https://openreview.net/forum?id=SkgKzh0cY7, 2019.
[14] Isola Phillip, Jun-Yan Zhu, Tinghui Zhou and Alexei A. Efros, Image-to-image translation with conditional adversarial networks, Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.
[15] Gao Chang, Derun Gu, Fangjun Zhang and Yizhou Yu, ReCoNet: Real-time Coherent Video Style Transfer Network, Asian Conference on Computer Vision, Springer, 2018.
[16] He Kaiming, Gkioxari Georgia, Dollar Piotr and Girshick Ross, Mask R-CNN, The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2961-2969.
[17] J Gauthier, Conditional generative adversarial nets for convolutional face generation, Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester, 2014(5):2, 2014.

© 2020 Information Processing Society of Japan