NTCIR-17 QA Lab-PoliInfo-4 Answer Verification における GDADC の利用に向けての考察

渋木英潔 [1]　内田ゆず [2]　小川泰弘 [3]　門脇一真 [4]　木村泰知 [5]
[1] 株式会社 BESNA 研究所　[2] 北海学園大学　[3] 名古屋大学
[4] 株式会社日本総合研究所　[5] 小樽商科大学
shib@besna.institute　yuzu@hgu.jp　yasuhiro@is.nagoya-u.ac.jp
kadowaki.kazuma@jri.co.jp　kimura@res.otaru-uc.ac.jp

概要

我々は，フェイクニュースなどの社会問題を質問応答や自動要約などの自然言語処理技術を用いて解決するためのシェアードタスクとして NTCIR-17 QA Lab-PoliInfo-4 (以下，PoliInfo-4) を開催している．本稿では，システムの改善に有効とされる敵対的データを継続的に収集するために，ゲーミフィケーションによるモチベーションで不特定多数のユーザに敵対的データを作成させる GDADC を説明する．また，GDADC のゲームとしての面白さを考察しながら，PoliInfo-4 のサブタスクである Answer Verification (以下，AV) での利用について述べる．

1　はじめに

政治にまつわるフェイクニュースが社会問題となって久しい．我々はフェイクニュースなどの社会問題を，質問応答や自動要約などの自然言語処理技術を用いて解決する取り組みとして，QA Lab-PoliInfo タスク [1, 2, 3] を評価型ワークショップ NTCIR[1)] において開催している．今年も NTCIR-17 QA Lab-PoliInfo-4 [4] (以下，PoliInfo-4) として，Question Answering-2 (以下，QA-2)，Answer Verification (以下, AV), Stance Classification-2, Minutes-to-Budget Linking の 4 つのシェアードタスクを開催している．

前回の QA Lab-PoliInfo-3 [3] で開催した Question Answering (以下，QA) では，例えば「○○について知事はどんな見解を持っているのか？」といった質問[2)]を入力として，一次情報となる知識源[3)]から

- 数値の誤り（特に年号）

質問	緑地整備を進めるべき。
Gold Standard	27年度早期に整備方針改定し、区市町との連携を更に深めながら整備を加速。
出力	2年度早期に整備方針を改定、区市町との連携を更に深めながら整備を加速させ、ゆとりと潤いのある東京の実現を図る。

- 要約元の選定失敗

質問	コロナで経済的格差が鮮明に。国と連携し生活底上げを。
Gold Standard	区市町村と連携し、各学校が現状に即した指導計画への再構築を行う。
出力	生活資金の無利子貸し付け等を講じている、支援を国の取組含め検索できるサイトを立ち上げ、情報が届く仕組みも整えている。

図 1　Question Answering における誤った解答例

質問の回答を提示した．参加者から提出されたシステムの回答を四つの観点から人手で評価した結果，800 点満点中 499 点 (62 %) が最高スコアとなり，適切な回答も多く出力されていた．しかしながら，Gold Standard[4)] を人手評価した結果の 598 点 (75 %) には及ばなかった．システムによる誤った回答の例を図 1 に示す．図 1 の上部に示す年号の誤りなど一見正しそうに見えるものもあるが，ファクトチェックの支援材料としてそのまま提示するには問題がある．それゆえ，PoliInfo-4 では，QA-2 として引き続き同じ設計のタスクを実施するとともに，質問応答システムの出力が質問に対して適切であるか否かを判断する AV を実施することとした．

AV は，QA などで生成された回答が知識源と照らし合わせてフェイクになっていないかをチェックするタスクである．QA と AV のイメージを図 2 にそれぞれ示す．AV は図 2 に示すように二値分類問題に帰結させることができる．したがって，近年の機械学習の進歩もあり，フェイクか否かのアノテーションが付与された十分な量の訓練データがあれば，適切な判断可能なモデルの生成が期待できる．しかしながら，QA の人手評価の対象は延べ 200 問

1）　https://research.nii.ac.jp/ntcir/index-ja.html
2）　『都議会だより (https://www.gikai.metro.tokyo.jp/newsletter/)』での代表質問や一般質問を用いた．
3）　『東京都議会会議録 (https://www.gikai.metro.tokyo.jp/record/)』を用いた．
4）　『都議会だより』の回答を用いた．

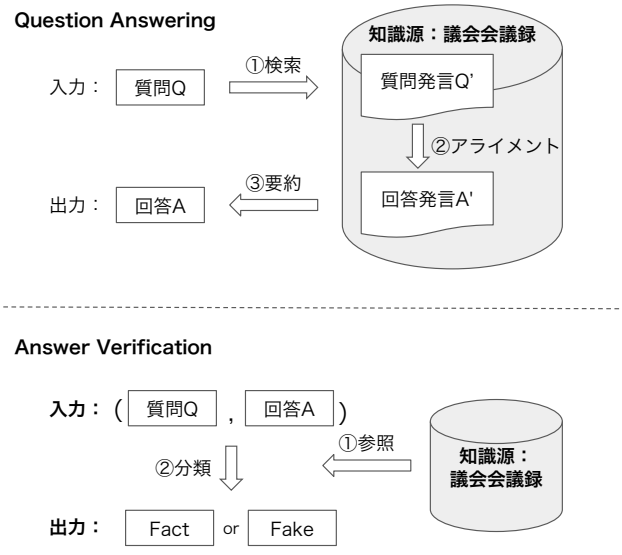

図 2 Question Answering (上) と Answer Verification (下) の概念図

程度であり，訓練データとして利用するには決して十分な量とは言えない．また，日本の政治または議員発言に関するフェイク情報という特殊性から，一般のタグ付きコーパス等を利用することも困難である．

この問題を解決するために，我々は，ゲーミフィケーション的にアノテーションデータを拡充する Gamified Dynamic Adversarial Data Collection (GDADC) を提案した [5]．本稿では，GDADC を説明した後，ゲームとしての面白さを考察しながら AV での利用に向けての取り組みを説明する．

2 ゲーミフィケーションに基づく動的な敵対的データ収集 (GDAGC)

近年，機械学習ライブラリの充実により，アノテーションデータさえあれば誰でも比較的容易に文書分類などのシステムを実装することが可能となった．アノテーション付きの言語資源は，言語資源協会[5]や ALAGIN[6]などから入手することができるが，その数量は限られており，目的とするドメインの言語資源がないこともある．質問応答 (QA) における機械読解を対象としたデータセットに SQuAD [6] があるが，Jia and Liang [7] はこれに人手で「システムが混乱するような敵対的な文 (adversarial distracting sentence)」を加えることでシステムの性能が低下することを示した．QA システムの精度改善のためにはシステムが間違いやすい部分に焦点を当てたデータが必要であるが，実世界の問題においてはそのようなデータが十分にあるとは限らず，特に専門性や特殊性が高い分野の場合，適格なアノテータを集めて新たにアノテーションすることも容易ではない．

専門知を補うために集合知を利用するという考え方は昔からあり [8]，クラウドソーシングを利用してコーパスを構築する研究も多く行われている [9]．Kiela et al. [10] は，クラウドソーシングを通して敵対的な事例を収集するために，アノテータが作成した事例に対してシステムが予測し，予測が誤った事例のみを収集する敵対的データ収集 (Adversarial Data Collection, ADC) を提案した．ADC により収集された事例は，予測システムが苦手とするタイプの事例に集中するのではないかという頑健性の問題が Kaushik et al. [11] により示されたが，Wallace et al. [12] は，敵対的データを元に予測システムを改善し，改善された予測システムを対象に新たな敵対的事例を収集するというサイクルを継続的に繰り返す「動的な敵対的データ収集 (Dynamic ADC, DADC)」により頑健なシステムが構築されることを示した．しかしながら，ADC に関する研究はクラウドソーシングが前提であり，Sugawara et al. [13] が示すようにアノテータの費用やインセンティブといった面で問題がある．従って，継続的にアノテーションを行うためには，インセンティブの問題を解決する必要がある．ユーザに自発的・持続的な行動を促すための研究として，ゲームに見られる様々な仕組みや要素をゲーム以外に適用するゲーミフィケーションと呼ばれるアプローチを適用したものがある [14, 15]．我々は，ゲーミフィケーションによりインセンティブを高めることで，不特定多数のユーザを対象として継続的にデータセットを拡充できると考えた．

以上の背景から，我々は**ゲーミフィケーションに基づく動的な敵対的データ収集** (Gamified DADC, **GDADC**) を提案した [5]．GDADC は，アノテーションの動機付けをゲームとしての面白さに頼っている．しかしながら，提案時にはその部分に関する議論が不十分であったため，本稿の 4 節で議論する．

3 Answer Verification での利用

AV では敵対的データ収集のために『AI 城から財宝を奪おう!』[7]というブラウザゲームを公開予定である．AV における GDADC を利用した敵対的デー

5） https://www.gsk.or.jp/
6） https://alaginrc.nict.go.jp/
7） https://sites.google.com/view/poliinfo4/game

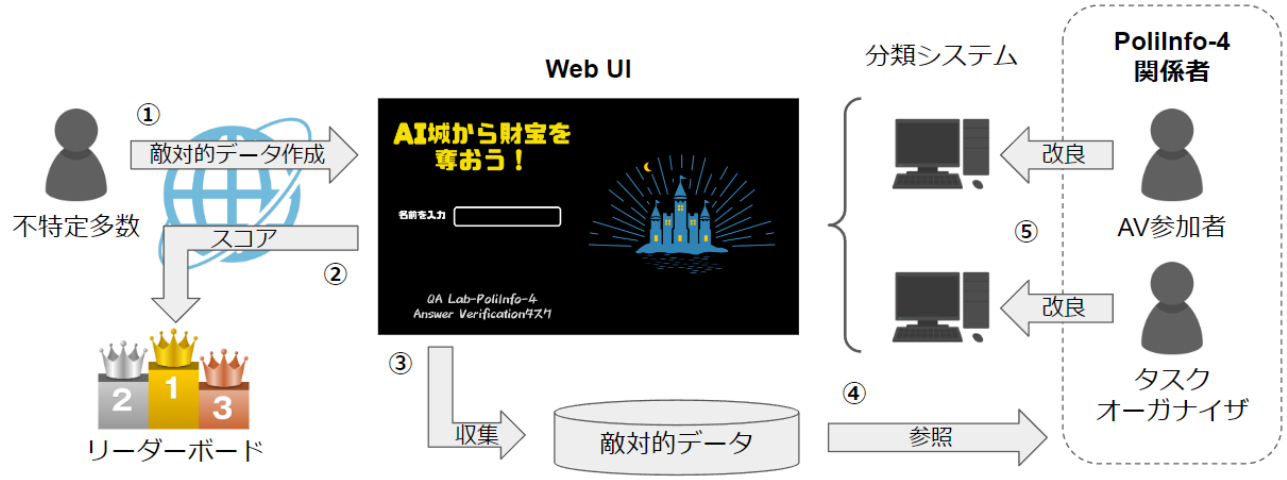

図 3　GDADC を利用した Answer Verification における敵対的データの収集

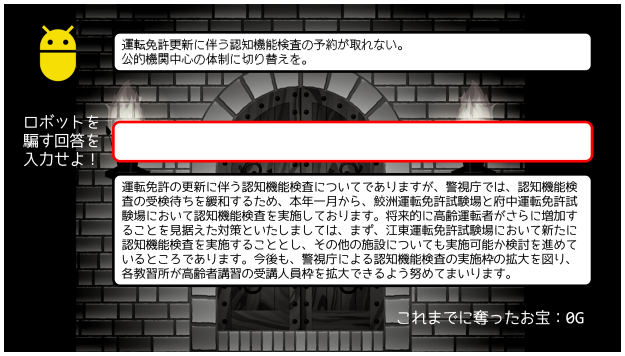

図 4　WebUI 上の敵対的データ作成画面 (予定)

タ収集の全体像を図 3 に示す．また，開発中の敵対的データの作成画面を図 4 に示す．GDADC のプレイヤは，敵対的データを作成する**攻略**と，予測システムを改良する**防衛**の一方または両方のアクションをとることができる．図 3 では，左側の Web 上にいる不特定多数が攻略側プレイヤ，右側の AV 参加者やタスクオーガナイザなどの PoliInfo-4 関係者が防衛側プレイヤに該当する．

ゲームストーリーとしては，攻略側プレイヤが「詐欺師」となり，防衛側プレイヤが作成した「AI 城主」から尋ねられた政治的質問に対して回答する (詐欺を働く) という形で，AI 城主が納得しつつも内容的には決して真実ではないテキストを入力する．その結果，AI 城主が真実だと予測すれば (すなわち，AI 城主が騙された場合)「そなたの意見はもっともだ．褒美をとらそう」となり，プレイヤの「財宝」スコアが増加する．スコアはリアルタイムで反映され，リーダーボードという形で順位が公開される．また，AI 城主を騙せたかどうかに関わらず，入力されたテキストはデータベースに収集される．防衛側と攻略側のデータ構造を表 1 と表 2 にそれぞれ

表 1　防衛側のデータ構造

フィールド名	説明
入力	
ID	識別番号
Meeting	会議名（都議会だよりの表記）
Date	日付（yyyy-mm-dd）
Headlines	質問者の発言全体の趣旨（2 文）
SubTopic	サブトピック
QuestionSpeaker	質問者
QuestionSummary	質問の要約
AnswerSpeaker	答弁者
AnswerSummary	答弁の要約
出力	
PredictedClass	AnswerSummary を fake と推測すれば “Fake”， fact と推測すれば “Fact”

示す．収集された敵対的データは，防衛側プレイヤに参照され予測システムの改良に用いられることになる．AI 城主が賢くなれば，同様の詐欺を働くことができなくなり，新たな種類のテキストを入力しなければならなくなる．このように攻略側プレイヤと防衛側プレイヤが互いに切磋琢磨することでゲームとしての駆け引きが生じることになる．

4　ゲームの面白さと利用動機の考察

AV 参加者やタスクオーガナイザといった防衛側プレイヤには，予測システムを改良するという目的が敵対的データ作成の動機となりうるが，攻略側プレイヤには，そのような動機はないため，ゲームとしての面白さをインセンティブにする必要がある．

4.1　ゲームの面白さ

ゲームの面白さとは何かを扱った研究として，藤江ら [16] や馬場 [17] の研究がある．彼らは，ゲー

表 2　攻略側のデータ構造

フィールド名	説明
入力	
ID	識別番号
Meeting	会議名（都議会だよりの表記）
Date	日付（yyyy-mm-dd）
Headlines	質問者の発言全体の趣旨（2 文）
SubTopic	サブトピック
QuestionSpeaker	質問者
QuestionSummary	質問の要約
SampleAnswer-Summary	fact のサンプルとなる答弁の要約（都議会だよりにおける記述）
AnswerSpeech	会議録における答弁
AnswerSpeaker	答弁者
出力	
GeneratedAnswer	答弁の要約
ArbitraryClass	作成者が GeneratedAnswer を fake だと思っていれば “Fake”，fact だと思っていれば “Fact”（WebUI では “Fake” 固定で非表示）
Reason	なぜ fake（fact）と思うかの理由（後で議論する際の参考情報，任意）
IsSubmission	スコア計算の対象とする場合は真，そうでなければ偽 WebUI では非表示（真に固定）
評価	
PollClass	プレイヤによる多数決で fake であれば “Fake”，fact であれば “Fact”（Answer Verification 参加者が投票）

ムの面白さを Csikszentmihalyi のフロー理論 [18] で言うところの最適覚醒と結び付け，ゲームの基本構成要素であるプレイヤ・ルール・ツール（インターフェイス）の三者のバランスが最適となったときにゲームの面白さが発生するとしている．完全には予測不可能なプレイヤという要素を含むとしつつも，最適覚醒につながる刺激として，新奇性と，不確定性や複雑性による情報負荷を挙げている．また，新奇性が失われた後も環境に対する統制感と能力を証明することによる効能感が繰り返し遊び続ける動機となりうるとしている．敵対的データを作成するという作業は一般的なアノテーションに比べて新奇性が高いものであり，「AI を騙せたら攻略成功」という攻略指標はそれなりの不確定性や複雑性が存在する．また，リアルタイムでスコアが反映される枠組みは統制感や効能感を与えてくれるだろう．

4.2　ゲーム利用動機

現代日本の大学生におけるゲーム利用動機を扱った研究には，井口 [19] がある．現代日本の大学生におけるゲーム利用動機には「空想」「承認」「趣向」「達成」「友達」「学習」「気晴らし」の 7 つの要因があり，「気晴らし」以外の動機が高いほどゲームへの没入度が高くなることが報告されている．「空想」は「現実とは違う世界で楽しむことができる」や「現実にはできないようなことができる」といった要素である．本稿の「詐欺師としてフェイクニュースを作成する」という設定は現実には行ってはいけないものであり，「空想」の動機付けにつながると考えられる．「承認」には「他人よりも上手なプレイヤになりたい」や「相手を負かすのが楽しい」といった競争の要素が含まれており，GDADC ではリーダーボードによるランキングがこの動機付けにあたる．「達成」は「課題を達成することが嬉しいから」や「遊んでいるうちに上達するのが楽しいから」といった要素であり，「学習」は「難しいことが理解できることがあるから」や「新しい知識を得ることができるから」といった要素であるが，AI を騙すことができるフェイクニュースを作成するという課題は，専門性や特殊性が求められる難しい課題と言ってよく，知識源となる議会での討論の内容を理解する過程は，多くの攻略側プレイヤにとって新しい知識を得ることになるだろう．残りの「趣向」は「絵や映像がきれいだから」や「音や音楽に惹かれるから」といった要素，「友達」は「友人と一緒に遊ぶのが楽しいから」や「友人との話題になるから」といった要素である．我々は「趣向」や「友達」に関する面白さにはあまり焦点を当てていないが，Web UI の演出や SNS などでの宣伝などを通して動機付けすることができればよいと考えている．

5　おわりに

本稿では，機械学習システムの改善に有効とされる敵対的データを継続的に収集するために，ゲーミフィケーションによるモチベーションで不特定多数のユーザに敵対的データを作成させる GDADC を説明した．また，GDADC のゲームとしての面白さを考察しながら，PoliInfo-4 の AV での利用について述べた．PoliInfo-4 の AV の開催とあわせて，今後『AI 城から財宝を奪おう!』を https://sites.google.com/view/poliinfo4/game で公開する予定である．また，ゲーム進行によりデータが蓄積されていく中で，GDADC の効果を検証する予定である．

謝辞

本研究は JSPS 科研費 21H03769，22H03901 の助成を受けたものである．

参考文献

[1] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Kotaro Sakamoto, Madoka Ishioroshi, Teruko Mitamura, Noriko Kando, Tatsunori Mori, Harumichi Yuasa, Satoshi Sekine, and Kentaro Inui. Overview of the NTCIR-14 QA Lab-PoliInfo task. In **Proceedings of the 14th NTCIR Conference**, 2019.

[2] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Teruko Mitamura, Masaharu Yoshioka, Tomoyoshi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Kenichi Yokote, Tatsunori Mori, Kenji Araki, Satoshi Sekine, and Noriko Kando. Overview of the NTCIR-15 QA Lab-PoliInfo-2 task. In **Proceedings of The 15th NTCIR Conference**, 2020.

[3] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Masaharu Yoshioka, Tomoyoshi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Kenichi Yokote, Kazuma Kadowaki, Tatsunori Mori, Kenji Araki, Teruko Mitamura, and Satoshi Sekine. Overview of the NTCIR-16 QA Lab-PoliInfo-3 task. In **Proceedings of The 16th NTCIR Conference**, 2022.

[4] 小川泰弘, 木村泰知, 渋木英潔, 乙武北斗, 内田ゆず, 高丸圭一, 門脇一真, 秋葉友良, 佐々木稔, 小林暁雄. NTCIR-17 QA Lab-PoliInfo-4 のタスク設計. 言語処理学会第 29 回年次大会, 2023.

[5] 渋木英潔, 内田ゆず, 小川泰弘, 門脇一真, 木村泰知. ゲーミフィケーションに基づく QA データセット拡充手法の提案: QA Lab-PoliInfo-4 Answer Verification タスクに向けて. 第 18 回 Web インテリジェンスとインタラクション研究会, 2022.

[6] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[7] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 2021–2031, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[8] Thatsanee Charoenporn, Virach Sornlertlamvanich, Hitoshi Isahara, and Kergrit Robkop. KUI: an ubiquitous tool for collective intelligence development. In **Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages**, 2008.

[9] 河原大輔, 町田雄一郎, 柴田知秀, 黒橋禎夫, 小林隼人, 颯々野学. 2 段階のクラウドソーシングによる談話関係タグ付きコーパスの構築. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2014, No. 12, pp. 1–7, 06 2014.

[10] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4110–4124, Online, June 2021. Association for Computational Linguistics.

[11] Divyansh Kaushik, Douwe Kiela, Zachary C. Lipton, and Wen-tau Yih. On the efficacy of adversarial data collection for question answering: Results from a large-scale randomized study. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 6618–6633, Online, August 2021. Association for Computational Linguistics.

[12] Eric Wallace, Adina Williams, Robin Jia, and Douwe Kiela. Analyzing dynamic adversarial training data in the limit. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 202–217, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[13] Saku Sugawara, Nikita Nangia, Alex Warstadt, and Samuel Bowman. What makes reading comprehension questions difficult? In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 6951–6971, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[14] 根本啓一, 高橋正道, 林直樹, 水谷美由起, 堀田竜士, 井上明人. ゲーミフィケーションを活用した自発的・持続的行動支援プラットフォームの試作と実践. 情報処理学会論文誌, Vol. 55, No. 6, pp. 1600–1613, 06 2014.

[15] 小河晴菜, 西川仁, 徳永健伸, 横野光. ゲーミフィケーションを用いたアノテーション付き対話データの収集基盤. 言語処理学会第 26 回年次大会, 2020.

[16] 藤江清隆, 馬場章. ゲームの面白さとは何か：テレビゲームのプレジャラビリティをめぐって. 日本バーチャルリアリティ学会誌 = Journal of the Virtual Reality Society of Japan, Vol. 9, No. 1, pp. 15–19, 03 2004.

[17] 馬場章. 2. ゲーム学の国際的動向. 映像情報メディア学会誌, Vol. 60, No. 4, pp. 491–494, 2006.

[18] Mihaly Csikszentmihalyi, 今村浩明. フロー体験喜びの現象学. Sekaishiso seminar. 世界思想社, 1996.

[19] 井口貴紀. 現代日本の大学生におけるゲームの利用と満足. 情報通信学会誌, Vol. 31, No. 2, pp. 67–76, 2013.