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

小河 晴菜[1]　西川 仁[1]　徳永 健伸[1]　横野 光[2]
[1] 東京工業大学 情報理工学院　[2] 株式会社富士通研究所
ogawa.h.ai@m.titech.ac.jp　{hitoshi,take}@c.titech.ac.jp
yokono.hikaru@fujitsu.com

1　はじめに

自然言語処理研究の様々な分野において，大規模なデータを用いてモデルを訓練する機械学習・深層学習によるアプローチが中心的である．これは対話研究においても例外ではなく，このようなデータ駆動アプローチにより訓練された高性能な対話モデルの構築が期待されている．このため，大規模な対話データが非常に重要となっている．

対話は，雑談のように目的を持たない非タスク指向型対話と，特定の目的を達成するために行うタスク指向型対話の二種類に分けることができる．このうち，非タスク指向型対話向けのデータにはインターネット上のチャットや映画の台本など既存のデータを利用できるが，タスク指向型対話向けには対話の目的に応じた状況を設定し実際に人間を用いてデータを構築する必要があるため，収集コストが高くなりデータ構築がより難しくなっている．

人手を必要とするデータ収集には，近年，クラウドソーシングが用いられることが多い．単純な作業を，多数の匿名作業者に低い単価で依頼するクラウドソーシングは，安価で早くデータを収集できることから，対話データを含む言語資源収集に用いられてきた [3, 4]．しかし，従来のクラウドソーシングによるタスク指向型データ収集には二つの問題点がある．まず一つに，作業者の動機付けの問題がある．クラウドソーシングによるタスク指向型対話データ収集では，作業者は目的を与えられ，それを達成するための対話を行う．この時，作業者の主な動機は目的達成ではなく作業報酬にあるため，いわゆる「やらせ」のような状況になり，収集したデータから自然さが失われてしまう危険性がある．二つ目の問題として，コストの増加がある．金銭を報酬とするクラウドソーシングでは，収集するデータの量に比例してコストが増える．大量のデータ収集には多くの作業者を必要とするため，作業者一人あたりの単価が低くても全体のコストは大きくなる．さらに，タスク指向型対話収集の場合，タグ付けのような単純な作業と違い目的達成までの拘束時間が長くなる．そのため，作業者あたりの単価も高くなる．

我々は効率的なタスク指向型対話データ収集を実現するために，ゲーミフィケーションを利用したデータ収集基盤を提案する．ゲーミフィケーションを用いることで，先に述べたクラウドソーシングにおける問題点を解決することを狙う．提案する基盤は，多様な対話タスクに利用可能な汎用性を持ち，データ収集者が求めるデータに応じて設定を変更･追加可能な拡張性を持つ．ゲーミフィケーションと汎用･拡張性を両立させるため，基盤には，Mojang 社が開発した Minecraft[1]（MC）を用いる．MC は三次元の仮想世界内を自由に探索・変更できるのが特徴のビデオゲームである．

さらに，我々は基盤上に，対話者が自分自身の発話をアノテーションする「発話者によるアノテーション [9]」を行う機能を構築する．対話データのアノテーションは対話と関係のない第三者によって行われることが多く，このアノテーション手法は今までにあまり注目されてこなかった．提案する基盤を通じてその可能性を探る．

2　ゲーミフィケーション

作業をゲーム形式で行うことで面白さによる動機付けを行うゲーミフィケーションは，教育や医療の他，自然言語処理研究にも適用されてきた [7, 8]．一方，汎用性を持ったタスク指向型対話データ収集基盤に適用した例はまだない．

この手法の大きな利点は，作業者の動機付けが自然な形でできる点にある．1 節で述べたように，クラウ

[1] https://www.minecraft.net/

ドソーシングでのタスク指向型対話データ収集では，必ずしも作業者が対話の目的を達成したいわけではないことから，不自然な対話となってしまう可能性がある．ゲーミフィケーションで面白さによる動機付けを行うことで，作業者が自発的に作業に取り組むようになり，自然な対話が行われることが期待できる．また，データ収集の際の金銭的コストの肥大化を抑えられるという利点もある．通常のクラウドソーシングでは，収集するデータ量に比例して作業者への報酬が増大する．一方，対話データを集めるためのゲームをインターネット上に公開し人を集めることができれば，主なコストはゲームの作成・運用コストのみになり，集めるデータ量に比例しない．他に，ゲーミフィケーションによりデータ収集タスクの認知負荷を軽減できる可能性がある．一般に，複数のことを同時に行うマルチタスクは認知負荷を増加させ，作業のパフォーマンスを低下させるが [6]，タスクにゲーミフィケーションを導入して実施した場合は認知負荷が減少する可能性 [5] が報告されている．

ゲーミフィケーションには上に挙げたような利点がある一方，作業者を動機づける魅力的なゲームを低コストで一から作成することは非常に難しい．この問題を避けるため，我々は既存のゲームである MC を利用する．以下に MC を選択した理由を述べる．

- MC を拡張して基盤を実装することで，新たにゲームを作成するよりも労力を抑えることができる．MC では mod と呼ばれるシステム拡張の文化が盛んであり，機能の拡張は比較的容易である．
- MC は世界中で一億人以上のアクティブプレイヤーが存在する，世界で最も人気なゲームの一つである．無名のゲームよりも，作業者を引きつけることが期待できる．
- MC には定められた目標がないため，データ収集者が構築したいデータに沿ったタスクを設定するのに都合が良い．

3 基盤

3.1 概要と使用例

我々が提案するデータ収集基盤は，データ収集者が自由に仮想世界を構築し，書き言葉での対話データを収集するための機能を提供する．基盤は特定のタスクに依存するものではなく，データ収集者は独自のタスクを設計し，実装することができる．また基盤は他に，作業者のペアを自動で作成する機能や，複数の作業者が並行してタスクを行えるよう分離された環境を用意する機能などを持つ．収集基盤の機能は，2 節で述べた MC の mod として実装されている．

提案する基盤で行うタスクの例として，我々は Map Task [1] を元にした Mansion Task を考案した．元となった Map Task は二人の作業者が違う役割を演じる非対称的なタスクであったが，Mansion Task では二人は対等な立場にあり，協力して共通の目的に挑む．作業者はそれぞれ内容が一部異なる屋敷の地図を渡される．地図には部屋の配置とその説明，ならびにスタート地点とゴール地点が記されている．このタスクの目的は，スタートからゴールまで辿り着くことである．ただし，ゴールに辿り着くためには,「鍵を見つけて扉を開ける」といったようなサブゴールをいくつか達成する必要がある．サブゴールを達成するための情報はどちらか一方の地図にしか書かれていないため，作業者らは対話によって情報を交換することになる．

提案した基盤をテストし，アノテーション付きのタスク指向対話データを収集するため，我々は基盤を用いて Mansion Task をゲーム化した．図 1 にそのスクリーンショットを示す．本来の Mansion Task は Map Task と同様ゲームと関係のないタスクであるが，基盤上では実際にプレイヤーが仮想世界内を歩き回りゴールへ向かうゲームとして実装した．追加要素として，1 節で述べた発話者によるアノテーションを実現するため，プレイヤーが自分たちの行った発話に対し対話行為のアノテーションを行うような機能を追加した．発話者は発話時に最も適した対話行為を選択する必要があり，もう一人のプレイヤーは任意に相手の発話をアノテーションできる．両方のプレイヤーがアノテーションを行うと，ゲーム上でスコアが付与される．プレイヤーのアノテーションに対する動機付けを高めるため，付与されるスコアはアノテーションの結果が一致した場合に高くなる．

3.2 発話者によるアノテーション

一般に，対話コーパスのアノテーションは，対話とは関係ない第三者によって行われることが多い．しかしながら，第三者は対話データから発話者の意図を推測する必要があり，誤った解釈により本来の意図とは違ったアノテーションを行う可能性がある．反対に，発話者自身がアノテーションをする場合には解釈誤りは起こりえないため，発話者がアノテーションのルールを理解していれば，原則誤ったラベルが付与されるこ

図 1: 基盤上での Mansion Task のスクリーンショット

ともない．これは発話者によるアノテーションの明確な利点であり，一つの手法として考慮する価値がある．

一方，発話者によるアノテーションは対話と同時にアノテーションを行うマルチタスクであり，作業者にかかる認知負荷が大きいという欠点がある．この欠点は，発話者によるアノテーションが一般に行われていない要因の一つであると推測する．ここで，2 節で述べたように，ゲーミフィケーションによって認知負荷を下げられる可能性が指摘されている．発話者によるアノテーションの有用性を探るため，発話者によるアノテーションを行う機能を基盤上に追加し，Mansion Task に組み込むことで欠点の緩和を狙う．

4 実験

4.1 実験方法

データ収集基盤のテストを行うと同時に，発話者によるアノテーションの有用性を検討するため，10 人（5 ペア）のプレイヤーによる小規模な実験を行った．プレイヤーらは 3.1 節で述べた Mansion Task を基盤上で行い，我々はペアごとの対話データ及びそのアノテーションを収集した．更に，集めた対話データそれぞれに対し，対話行為のアノテーション経験がある二人のアノテータによるアノテーションを行なった．つまり，一つの発話に対して，二人のプレイヤーと二人のアノテーション経験者，計四人分のラベルが付与される．

五ペアによる五つの対話について，各アノテータ間での Cohen のカッパ係数 [2] ならびに各アノテータの正解率を計算した．ここで，対話内では，各プレイヤーは話し手と聞き手の二つの役割を演じる．話し手は発話の意図を理解しているという前提から，アノテーションの正解率は，発話者のラベルを正解として計算する．また，聞き手がアノテーションを行うかどうかは任意であるため，正解率の算出には聞き手がラベルを付与した発話のみを用いて算出した．

4.2 実験結果

表 1 にペアごとの Cohen のカッパ係数を示す．一方表 2 はアノテーションの正解率を示す．太字はその行での最大値を表している．正解率がすべて 0 の行が存在しているのは，その対話で聞き手がアノテーションを行ったのが一発話のみだったことが原因である．表から，Cohen のカッパ係数ではアノテーション経験者間の値がプレイヤーを含むペアより比較的高い一方で，正解率はアノテータによってばらつきがあり，10 対話中 5 つでアノテーション経験者の正解率がプレイヤーを下回ったことが分かる．カッパ係数と正解率の値に差があることは，アノテーション経験者が合意したラベルが，必ずしも発話者の意図を反映したものではないということを意味している．

表 1: 五つの対話でのアノテータ間のカッパ係数
(P1・P2: プレイヤー, A1・A2: アノテーション経験者)

		P1	P2	A1
1	P2	0.752		
	A1	0.725	0.697	
	A2	0.767	0.683	0.889
2	P2	0.530		
	A1	0.627	0.677	
	A2	0.627	0.579	0.668
3	P2	0.140		
	A1	0.159	0.468	
	A2	0.175	0.552	0.764
4	P2	0.318		
	A1	0.362	0.512	
	A2	0.363	0.491	0.901
5	P2	0.595		
	A1	0.574	0.710	
	A2	0.574	0.645	0.908

この結果は，たとえ対話行為アノテーションの経験があっても，第三者であるアノテータにとって発話者の意図を推測するのが難しい場合があることを示唆している．例を挙げると，ある対話の「学生室ってところに青の鍵がかかった宝箱があるらしいけども」という発話に対し，プレイヤー二人は「提案」というラベルを付与したが，アノテーション経験者二人は「伝達」を付与していた．この時，発話者は学生室に行くこと

表 2: 各発話者に対するアノテータの正解率
（H: 聞き手, A1・A2: アノテーション経験者）

Dialogue	H=P1	A1	A2
1	**0.870**	0.804	0.826
2	0.792	**0.849**	0.792
3	0.000	0.000	0.000
4	**0.696**	0.609	0.609
5	0.750	**0.813**	**0.813**
Dialogue	H=P2	A1	A2
1	0.744	0.721	**0.767**
2	0.649	0.688	**0.714**
3	**0.720**	0.520	0.600
4	**0.821**	0.679	0.643
5	**0.808**	0.731	0.731

を提案するためにこの発話を行い，もう一人のプレイヤーもその意図を正しく認識した一方，アノテーション経験者らはただの情報伝達だと受け取ったと考えられる．発話者の意図を正しく認識するためには，状況や雰囲気，ならびに話し手の行動のような非言語情報が必要な場合があるが，一般に第三者であるアノテータらはこれらの情報を手に入れることが難しい．一方，この例における聞き手は，第三者と同じように話し手の意図を想像する必要があったものの，状況が共有されていることによりアノテーションに成功したと推測する．

このように，言語情報以外が重要な役割を担うような対話では，第三者のアノテータが発話意図の解釈誤りを起こす場合がある．このような対話において，発話者によるアノテーションは一定の有用性があると考える．

5 おわりに

本論文では，ゲーミフィケーションを用いたタスク指向型対話データ収集基盤を提案した．提案した基盤は，クラウドソーシングによる対話データ収集が持つ課題を，ゲーミフィケーションによって解決することを狙っている．多様なタスクに対応できる汎用性を持つ基盤を構築するため，Minecraft を用いて実装を行った．また，基盤上で発話者によるアノテーションを行うことを提案した．発話者によるアノテーションにゲーミフィケーションを導入し，その可能性を探った．

提案したデータ収集基盤上で小規模なデータ収集実験を行った．実験には，Map Task を元に我々が考案した探索タスクである Mansion Task を用いた．このゲームを通じて対話データを収集し分析を行った結果，第三者のアノテータは，発話者の意図を正確に汲み取れない場合があることや，そのような場合での発話者によるアノテーションの有用性が示唆された．今後はさらにシステムの改良を行い，より大きな規模での収集実験によって収集基盤の可能性を探る．基盤は改修後に公開する予定である．

謝辞

本研究は JSPS 科研費 JP19H04167 の助成を受けたものです．

参考文献

[1] Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. The HCRC map task corpus. *Language and Speech*, Vol. 34, No. 4, pp. 351–366, 1991.

[2] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37–46, 1960.

[3] Walter Lasecki, Ece Kamar, and Dan Bohus. Conversations in the crowd: Collecting data for task-oriented dialog learning. In *In Proceedings of the Human Computation Workshop on Scaling Speech and Language Understanding and Dialog through Crowdsourcing at HCOMP 2013.*, January 2013.

[4] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 254–263, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[5] Chung-Ho Su. The effects of students' motivation, cognitive load and learning anxiety in gamification software engineering education: a structural equation modeling study. *Multimedia Tools and Applications*, Vol. 75, No. 16, pp. 10013–10036, August 2016.

[6] John Sweller. Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, Vol. 4, No. 4, pp. 295 – 312, 1994.

[7] Daniele Vannella, David Jurgens, Daniele Scarfini, Domenico Toscani, and Roberto Navigli. Validating and extending semantic knowledge bases using video games with a purpose. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1294–1304, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[8] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pp. 319–326, New York, NY, USA, 2004. ACM.

[9] 松吉俊．反転学習とシステム開発演習を活用するテキストアノテーション．情報処理学会研究報告，Vol. 2017-SLP-116, No. 7, pp. 1–9, May 2017.