フレーミングに基づいた協調的説得対話方策の強化学習 *

◎平岡 拓也, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲 (NAIST)

1 はじめに

近年，説得・交渉対話に強化学習を適用した研究が行われている [1]．本論文では，フレーミング（2節）を用いた協調的説得対話システムの方策の強化学習に取り組む.「協調的説得対話」とは，説得をしながら，被説得者の目標も満足しようとする対話である．強化学習を適用するため，人同士の説得対話コーパス（2節）を利用し，部分観測マルコフ決定過程（POMDP）（3，4節）を構築する．そして，学習された方策の性能を評価 (5節) を行う．

2 説得対話コーパスとフレーミング

本研究では，POMDP 構築に，説得対話コーパス [2] を利用する．本コーパスは，説得対話の一例として，家電販売店でのカメラ販売における販売員（説得者）と客（被説得者）の対話を想定する．販売員は客に対して，複数のカメラ（意思決定候補）の中から特定のカメラ (説得目標) を購入（意思決定）させることを目的とする．合計 35 対話（340 分）の模擬対話コーパスを後節のモデル構築に利用する．

コーパスには，ネガティブ/ポジティブ**フレーミング** [3] が注釈される．これらのフレーミングでは，感情極性を持った語で意思決定候補を修飾する．具体的には，ネガティブフレーミングはネガティブな感情極性を，ポジティブフレーミングはポジティブな感情極性を持つ語で意思決定候補を修飾する．本コーパスでは，フレーミングは 3 つ組 $\langle a, p, r\rangle$ で表される．a_i は論証の対象である意思決定候補を表す．p_i はフレーミングがネガティブの場合 NEG，ポジティブの場合 POS の値をとる．r_i は論証中に被説得者の嗜好に合致した決定要因（例 : カメラの性能や値段）への言及が存在するかを表す．r_i は言及が存在する場合 TRUE，存在しない場合 FALSE の値をとる．被説得者の嗜好に合致する決定要因はアンケート結果に基づいて決定する．表 1 はフレーミングの例である．

また，一般的な対話行為（例 : 質問や情報提示）として，一般目的機能（GPF）[4] も注釈する．

3 ユーザシミュレータ

強化学習時の報酬計算のため，ユーザ（2節の被説得者）の以下の振る舞いのシミュレータを構築する :

1. ユーザの一般的な対話行為.
2. ユーザへの嗜好の通知.

ユーザの一般的対話行為は GPF を用いて表わされる．また，ユーザへの嗜好の通知とは，説得者が代替案のフレーミングに引用した決定要因が被説得者の嗜好に合致することである．例えば，表 1 では，店員のカメラAのポジティブフレーミングに “性能” が引用されている．もし，“性能” が被説得者の好みに合致 (i.e. `pref=YES`) する場合は嗜好の通知がされたとする．

Table 1 フレーミングの例. $\langle a_i = \mathrm{A}, p_i = \mathrm{POS}, r_i = \mathrm{NO}\rangle$. 本例では，客の嗜好はカメラの値段にある．

(カメラAは) ポケットに入る大きさで一眼並みの性能で撮っていただけるっていうことが今回のポイントなんですけれども

ターン T_{t+1} における，ユーザの GPF G^{t+1}_{user} と嗜好の通知 C^{t+1}_{alt} はそれぞれ以下の式に基づいて計算される．

$$P(G^{t+1}_{user}|G^{t}_{user}, F^{t}_{sys}, G^{t}_{sys}, S_{alt}) \quad (1)$$

$$P(C^{t+1}_{alt}|C^{t}_{alt}, F^{t}_{sys}, G^{t}_{sys}, S_{alt}) \quad (2)$$

G^{t}_{Sys} はターン T_t におけるシステムの GPF，F^{t}_{Sys} はターン T_t におけるシステムのフレーミングを表す．これらはいずれもシステムのアクションであり，4節で説明する．G^{t}_{user} はターン T_t におけるユーザの GPF，C^{t}_{alt} はターン T_t における嗜好の通知状態を表す．S_{alt} は代替案の初期選択である．本研究では，ユーザが最初に嗜好に合致するとして選んだカメラである．

4 協調的な説得対話方策の学習

本節では，システム（2節の説得者）に関するモデルについて述べる．特に，強化学習を行う上で必要な情報である報酬や，システムの行動と信念状態について説明する．

我々はユーザの満足度（被説得者の目標の達成度合い），システムの説得成功（説得者の目標の達成度合い）と自然性を用いて報酬を設計する．1節で述べたように，我々は協調的な説得対話システムの構築を目指してる．従って，システムはユーザとシステム両方の目的を達成するよう対話を進めなければならない．各ターンにおける報酬の計算式は以下のとおりである．

$$r_t = (Sat^{t}_{user} + PS^{t}_{sys} + N^{t})/3 \quad (3)$$

Sat^{t}_{user} は，[0,1] に規格化された，ターン t における 5 段階のユーザの満足度の主観評価値の (1: Not satisfied，3: Neutral，5: Satisfied) を表す．PS^{t}_{sys} はターン t における説得の成功 (1: Success，0: Failure) の期待値である．N_t はターン t におけるシステムとユーザの対話の bi-gram 尤度である．なお，Sat^{t}_{user} と PS^{t}_{sys} は，先行研究 [2] で得られた予測モデルに基づき，対話状態（表 2）を利用して計算される．

システムのアクションはフィルタリングされたフレーミングと GPF の組 $\langle F_{sys}, G_{sys}\rangle$ である．これらは 2 節で述べた店員（説得者）の対話行為を表す．フィルタリングのために，実対話コーパスから店員のユニグラム $P(G_{sales}, F_{sales})$ を構築する．本研究では，$P(G_{sales}, F_{sales})$ が 0.005 以下の $\langle G_{sys}, F_{sys}\rangle$ を削除し，残った 13 個組をアクションとして利用する．

システムの信念状態は，報酬計算に用いた特徴量 (表 2) と報酬で表わされる．ただし，本研究では，システムは C_{alt} を観測できないと仮定し，式 (2) を用いて計算された推定値を利用する．

5 実験的評価

本節では，ユーザシミュレータと実際のユーザに対する，フレーミングと学習した方策の有効性検証を目的とした実験を行う．

---

* Reinforcement Learning of Cooperative Persuasive Dialogue Policies using Framing . by Takuya Hiraoka, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura (NAIST)

Table 2　報酬計算のための特徴量

Sat_{user}	システムの commisive（GPF の一種）の頻度
	システムの question（GPF の一種）の頻度
PS_{sys}	経過時間
	ユーザへの嗜好の通知 C_{alt}
	ユーザの代替案の初期選択 S_{alt}
N	システムとユーザの現在のターンの GPF
	システムとユーザの直前のターンの GPF
	システムのフレーミング

5.1　方策学習とユーザシミュレータ評価

ユーザシミュレータに対して，フレーミング及び学習効果を検証するために以下の 3 つの方策を用いる．

Random ベースラインその 1．全てのアクションから当確率でランダムにひとつのアクションが出力される．

NoFraming ベースラインその 2．フレーミングを含まないアクションのみを用いて学習された方策に基づいてアクションが出力される．

Framing 提案手法．全てのアクションを利用して学習された方策に基づいてアクションが出力される．

評価のために Neural fitted Q Iteration[5] を用いて方策の学習を行う．学習では，50 対話を 1 セットとして，各セットごとに価値関数のパラメータの更新を行う．学習は 200 セット行い，全セットの中で最も高い報酬を獲得したの方策を評価用の方策とする．そして，ユーザシミュレータを用いた対話における 1000 対話の平均報酬に基づいてシステムの評価を行う．

評価結果（図 1）から，1) 学習により性能が向上し，2) フレーミングが有効であることの 2 点が示唆される．

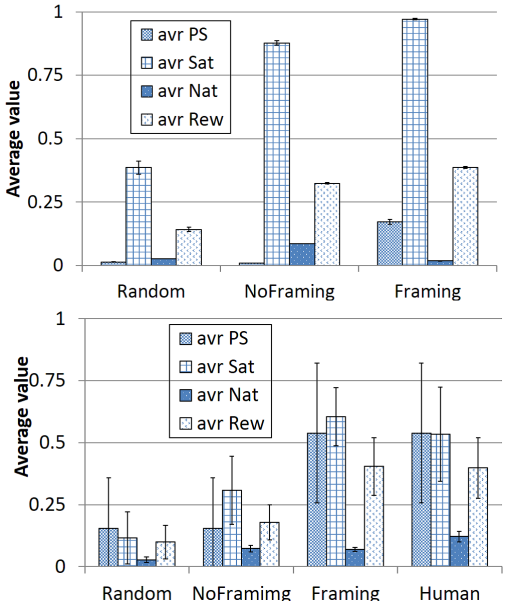

Fig. 1　各方策の平均報酬（上図：ユーザシミュレータ　下図：実ユーザ）．エラーバーは 95%の信頼区間を表す．Rew は報酬，Sat はユーザ満足度，PS は説得成功率，Nat は自然性をそれぞれ表す．

5.2　Wizard of Oz 法に基づいた実ユーザ評価

実際のユーザに対して，フレーミング及び前節の学習効果を検証する．本節では，5.1 節の方策に加え，以下の方策も評価する．

Human 人間（カメラ販売における説得対話の分析歴約 1 年）がアクションを選択する．

実験的評価は Wizard of Oz の枠組みに基づいて行われる．この評価では，システムは販売員を，評価者は客をそれぞれ演じる．システムの音声認識・言語理解，と言語生成は Wizard により行われる．音声認識・言語理解として，Wizard は評価者の発話を聞き，適切な GPFG_{user} に変換して，方策部に引き渡す．また，言語生成として，Wizard は類似発話に基づき，システムの応答文を作成し，テキスト音声合成部に引き渡す．類似発話とは，これまでの対話におけるシステムのアクションと G_{user} の系列に合致する説得対話コーパス中の店員の発話である．対話の最後には，評価者は，4 節の報酬を計算するために必要な情報を質問用紙に記入する．

実験参加者は評価者 13 人 (女性 3 人，男性 10 人) であり，各方策に従う Wizard とそれぞれ 1 回ずつ対話（計 4 対話）を行う．

実験結果 (図 1) から，フレーミングは実際のユーザに対しても有効であることが示唆される．なぜなら，Framing に対する評価は NoFraming と Random に比べて高く，Human と同等だからである．なお，NoFraming については Random とほぼ同等の評価を得ており，実際のユーザに対しては有効な方策でないことが示唆された．

Framing の方策に着目して考察すると，先行研究 [2] での，人間の対話における特徴をいくつか再現していることが分かった．よく見られた特徴のひとつは，説得目標であるカメラ A のポジティブフレーミングを行う際，カメラ B についても薦めることである．この特徴は，人同士のカメラ販売対話における説得が成功した場合に，よく見られた．

6　結論

本研究では，フレーミングを用いた協調的な説得対話システムの方策を強化学習した．強化学習を適用するため，説得対話コーパスを用いて，ユーザシミュレータと報酬関数を構築した．そして，学習された方策とフレーミングの効果を検証するため，ユーザシミュレータと実ユーザに対して性能評価実験を行った．評価実験から，強化学習を適用することはフレーミングを用いた協調的な説得対話システムに有効に働くことが示唆された．今後の予定として，音声認識・言語理解部と言語生成部を備えた説得対話システムを構築することがあげられる．

参考文献

[1] K. Georgila, “Reinforcement learning of two-issue negotiation dialogue policies,” *Proc. SIGDIAL*, 2013.

[2] T. Hiraoka and et al., “Construction and analysis of a persuasive dialogue corpus,” *Proc. IWSDS*, 2014.

[3] L. Irwin and et al., “All frames are not created equal: A typology and critical analysis of framing effects,” *Organizational behavior and human decision processes 76.2*, 2013.

[4] ISO24617-2, *Language resource management-Semantic annotation frame work (SemAF), Part2: Dialogue acts. ISO*, 2010.

[5] M. Riedmiller, “Neural fitted Q iteration - first experiences with a data efficient neural reinforcement learning method,” *Machine Learning: ECML*, 2005.