雑談対話システムにおける強化学習を用いた応答生成モジュールの選択

江頭 勇佑　　柴田 知秀　　黒橋 禎夫
京都大学大学院情報学研究科
egashira@nlp.ist.i.kyoto-u.ac.jp, {shibata, kuro}@i.kyoto-u.ac.jp

1　はじめに

観光案内や切符の予約システムといった，ある特定のタスクを遂行する課題遂行型の対話システムはこれまでにも多く研究されており，少しずつ実用化もされ始めている [5]．その一方で，高齢者や子どもの話し相手 [4]，あるいはエンターテイメントとして，話すこと自体を目的とした雑談対話を行う非課題遂行型の対話システムの必要性も高まっている．

本研究ではユーザにとって未知の情報を提供する雑談対話を行うために，Web 上のニュース記事や Wikipedia などの知識，および質問応答などの技術を用いた応答生成モジュールを複数用意し，ユーザ入力に対してそれらから適切なものを選択することで対話を実現するシステムを構築する．多様な対話を行えるシステムを構築するためには複雑な選択規則が必要となり，そのような対話戦略を人手で記述することは難しい．そこで本研究では，強化学習の枠組みであるマルコフ決定過程（MDP）を用いて過去の対話から最適な対話戦略を学習する．またユーザとシステムとの実際の対話から学習を行うため，個別のユーザに適応した対話戦略を獲得できるという利点も存在する．

2　関連研究

Levin らは飛行機の対話型フライト案内システムの制御に MDP を用い，「どのようなユーザ発話に対してどのようなシステム発話をとるか」といった対話戦略について最適なものを自動獲得することに成功している [1]．

また，Young らは MDP に隠れ状態を適用し曖昧性を考慮できるように拡張した部分観測マルコフ決定過程（POMDP）を対話制御に用いることで，認識誤りの存在する音声対話システムにおいてより高い精度で対話を行うことができると示している [3]．

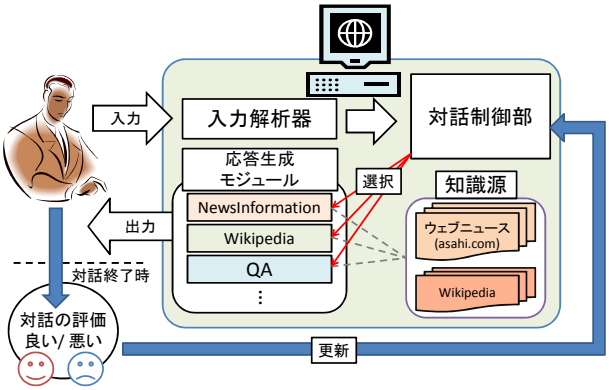

図 1　システムの概要

Meguro らはこの POMDP をシステムが聞き役となる雑談対話に利用し，学習によって獲得した対話戦略が雑談対話にも有用なものであることを示している [2]．しかし Meguro らの研究では生成された対話行為タグの列のみを扱っており，実際にユーザとシステムが対話を行う本研究とは異なる．本研究では音声入力を扱わず，ユーザの意図の曖昧さも一つの状態として扱うため，POMDP ではなく MDP の枠組みを用いる．

3　システムの概要

本システムでは，通常の発話生成や質問応答などの各機能を個別の応答生成モジュールとして扱い，ユーザ発話に対してそれらを適切に選択することで雑談対話を実現する．それぞれの機能を独立な応答生成モジュールとして構築することで機能追加や修正の利便性を確保できる．また，ユーザにとって未知の情報を提示するために，発話を生成するための主な情報源として Web 上の最新のニュース記事を用いる．例えば「あるイベントが開催される」という情報についての日時や場所といった情報や固有名詞の定義に関する質問に，質問応答や Wikipedia の知識を用いてシステムが回答する

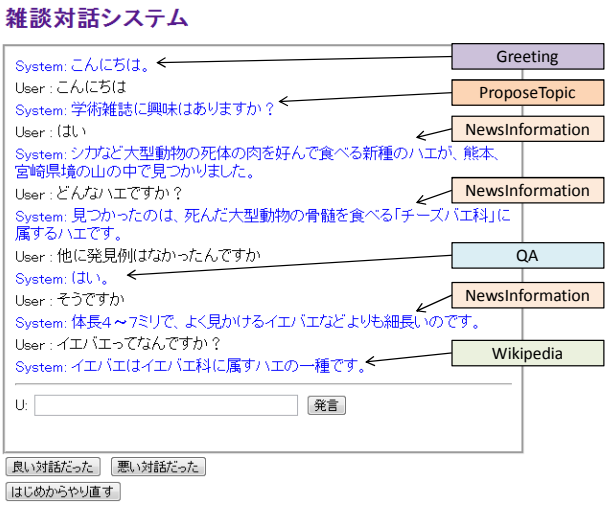

図 2　対話例

ことで，雑談的な対話を行いつつユーザに情報を提供することができる．

システムの概要を図 1 に，対話例を図 2 に示す．システムはユーザの発話を入力として受け取り，入力解析器においてそれがどのような種類の入力であるかを判断する．そして各応答生成モジュールにおいて入力に応じた発話が生成できるかどうかを検査し，それらの結果を総合したシステムの状態に対して，最適なシステム発話を出力できる応答生成モジュールを選択する．

対話を終了する際にユーザはその対話が良かったか否かの評価を行い，システムはその評価と対話の長さをもとにして対話戦略を更新する．強化学習の特性により，良い評価が与えられればその対話で選択されたモジュールが選択されやすくなり，逆に悪い評価であれば選択されにくくなる．このようにして次回からの対話における対話戦略が改善される．

4　応答生成モジュール

本システムで用いる応答生成モジュールを表 1 に示す．システムはあらかじめ獲得しておいたニュース記事とその解析を行ったデータ，Wikipedia の定義文データベースなどを保持しており，各応答生成モジュールではこれらの知識をもとに文の検索や質問応答などの処理を用いてシステム発話を生成する．各応答生成モジュールは相互に依存せず独立にシステム発話を生成するため，修正や追加が容易である．また，入力に対してあまりにも的はずれな発話を行わないように，例えば質問応答モジュールでは入力が疑問文でなければ発話を生成しないといった判断を，各応答生成モジュールで行っている．

表 1　応答生成モジュール一覧

応答生成モジュール	機能
Greeting	挨拶を行う
NewsInformation	ニュース記事中の一文を表示する
ProposeTopic	システムから話題を提案する
SearchTopic	ユーザが希望する話題の検索を行う
Wikipedia	Wikipedia 定義文の検索を行う
QA	ユーザの質問に対し回答を提示する
AskUser	ユーザに問いかける

知識として用いるニュース記事は可能な限り新しいものを利用するために，Web 上のニュースサイトより一定時間ごとに取得し，発話生成に必要となる形態素解析，構文解析，固有表現解析などの処理を行う．また各記事にはその内容をよく示しているキーワードを設定し，記事を話題として提案する際に利用する．

5　強化学習による対話戦略の獲得

ユーザとの対話を行ううえで重要になるのは，どのような入力に対しどのような出力を行うかという対話戦略である．従来の対話システムにおいては人手でルールを記述することが一般的であった．しかし「日本シリーズについて教えてください」という発話のように“日本シリーズの結果が知りたい”あるいは“日本シリーズという語句そのものの定義を知りたい”といった複数の意図が考えられる場合において，文脈などの状況に対応したルールを全て記述するのは大変である．ユーザごとに発話に込めた意図や期待する応答が異なっていることも考えられる．また，記述されるべきルールは応答生成モジュールに強く依存するため，機能の追加や修正に伴うルール修正に多大な労力を必要とするといった問題点も存在する．

そこで本研究ではこれらの問題に対処するため，対話戦略をシステムの状態に対する応答生成モジュール選択の問題とみなし，強化学習により自動獲得する．

5.1　MDP

強化学習では環境の状態 s を観測し，それに対応した行動 a をとって報酬 r を獲得するといったサイクルを繰り返す．そして 1 回の試行，ここでは一連の対話における獲得報酬を将来的に最大化するように対話戦略 π を更新することで最適な戦略を獲得する．本研究ではある応答生成モジュールを選択することを行動 a と定義する．

強化学習においてマルコフ性が満たされていると仮定した場合を MDP と呼び，その状況下では時刻 $t+1$

における環境の状態 s_{t+1} は時刻 t における状態 s_t とシステムによる応答生成モジュールの選択 a_t のみに依存することになる．対話に関しては，長い文脈などを考慮するとマルコフ性が完全に満たされていると言えないが，本研究では簡単のために MDP の枠組みで対話戦略を考える．

5.2 状態

本研究では，時刻 t にシステムが観測する状態 s_t を $\{UserInput, PreviousAction, ActionCandidate\}$ の 3 つの素性の組として扱う．

■UserInput　*UserInput* はユーザの入力を表 2 に示す属性の集合で表したものである．従来研究では対話制御部において発話に相当する各文はそれぞれ一つの対話行為タグに関連付けがなされていた．しかし雑談対話上では発話の意図について曖昧性を含む発話文を 1 個のタグに関連付けることは難しい．また，そのような発話文の持つ曖昧性を含めてシステムの取るべき行動を判断することで柔軟な対話を行うことができるため，本研究ではこのような属性の集合として表現する手法を取る．

■PreviousAction　*PreviousAction* はシステムが直前に行った選択である．この値により，マルコフ性を満たしつつも仮想的に直前より前の文脈を考慮することができる．

■ActionCandidate　応答生成モジュールが発話を生成するかどうかの判断を対話制御部とは独立に行う本システムにおいては，同じ *UserInput* で表現されるユーザ発話に対しても，どの応答生成モジュールで発話が生成可能かという点でいくつかの組合せが考えられる．そのため各応答生成モジュールが発話を生成できたかどうかというフラグの集合である *ActionCandidate* を状態に加える．これにより複数の応答生成モジュールにおいて発話の生成が可能である場合に，どれを優先的に選択するか学習できる．

5.3 報酬

時刻 t においてシステムが応答生成モジュールの選択 a_t を行ったことで得られる報酬として，まず各時刻 t ごとに $+1$ の報酬を設定する．これはシステムが少なくともユーザが許容できる発話を行うことができたと判断するためである．

対話終了時にユーザから良い対話であったと評価された場合には，満足できる対話を行うことができたとして $+20$ の報酬を得る．一方で対話が悪評価であった場合には，ユーザに対して不適切な反応をとる選択を行ったと判断し，特別な報酬は得られない．

表 2　UserInput の属性一覧

属性	内容
Declarative	平叙文
Question	疑問文
ChangeOtherTopic	話題の変更を求める
SearchOtherTopic	話題の検索を求める
Positive	肯定
Negative	否定

5.4 学習アルゴリズム

本研究では状態 s の遷移確率 $P(s_{t+1}|s_t, a_t)$ を定義することが難しいため，行動価値 $Q(s, a)$ に対するモンテカルロ法を用いて対話戦略 π を学習する．行動価値とは状態 s を観測したシステムが応答生成モジュールの選択 a を行うことで将来的に得られる報酬の期待値である．このアルゴリズムでは任意の対話戦略 π を用いて一度対話を行い，その結果得られた報酬をもとに行動価値 $Q(s, a)$ を更新，さらにその $Q(s, a)$ の値から新たな対話戦略 π を獲得する．

$Q(s, a)$ の値は，各時刻 t で状態 s_t に対し行った選択 a_t に対する将来的な報酬 R_t の期待値として以下のように計算する．

$$Q(s, a) = E\left(R_t | s_t = s, a_t = a\right)$$

R_t は時刻 t 以降に得られた報酬であり，次の式で表される．

$$R_t = \sum_{k=0}^{T-t} \gamma^k r_{t+k}$$

ここで r_t はターン t において獲得した報酬で，T は対話が終了した時刻である．これらの値は対話の履歴から容易に知ることができる．γ は未来に獲得されるであろう報酬の不確定さを導入するための割引率であり，本研究では $\gamma = 0.9$ とする．

行動価値 Q の初期値 Q_0 は 10 とする．本研究の報酬設定では，時刻 t が対話の終了時刻 T よりも遠くなるほど R_t の値は 10 に収束する．そのため，評価が良い対話の中で行われた選択は次回以降優先的に選択され，逆に悪い対話の中で行われた選択は選択されにくくなるという動作を効果的に実現できる．

6 実験

実験参加者 4 名がそれぞれ初期値からの対話戦略を用いて本システムと対話を行い，各 50 回の対話データを収集した．n 回目の対話までの，4 名のユーザ発話数

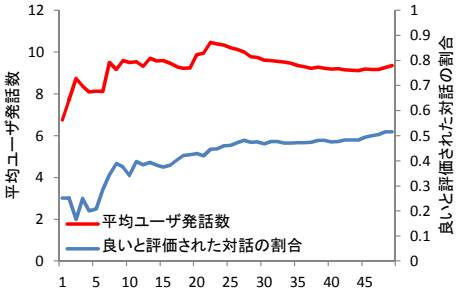

図 3　n 回目の対話までの平均ユーザ発話数と良評価の割合

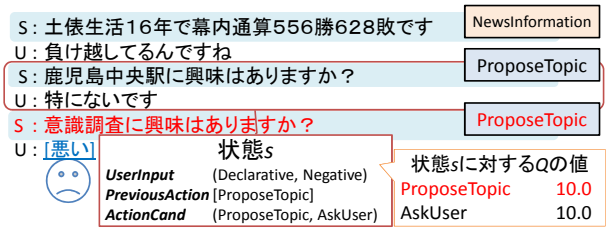

図 4　4 回目の対話における選択

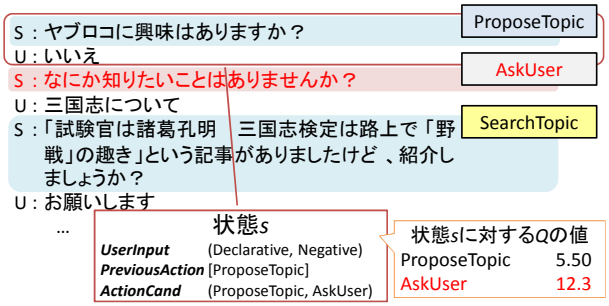

図 5　39 回目の対話における選択

の平均をとったものと，良いと評価された対話の割合を図 3 に示す．このグラフにより，対話を重ねることによりユーザの満足のいく対話を行うことができるような対話戦略を学習できていることがわかる．また平均ユーザ発話数もゆるやかに増加し，対話を長く続けられる戦略を獲得している．最終的には 9 回程度の発話数に収束しているが，これは発話の主な情報源であるニュースの記事が 10 文程度で構成されており，対話を行ったユーザが長くとも話の区切りの良いところで対話を終了するためにこのような結果になっていると思われる．

また，実際に適切な戦略を学習できている例を図 4 と図 5 により示す．図 4 はあるユーザの 4 回目の対話の一部である．図 4 の状態 s にあるときのシステムの選択 *ProposeTopic* に対してこのユーザは不適切な発話であると感じ，悪い対話であったとして終了した．その後 39 回目の対話である図 5 においても同じ状態 s が出現したが，それまでの対話から *AskUser* を選択する方がこのユーザに対して有効であると学習できており，ユーザもそのまま対話を継続した．一方で別のユーザによって学習された対話戦略においては，同じ状態 s に対して *ProposeTopic* を選択する方が好ましいと学習されていた．このような例から，ユーザに応じて適切な応答生成モジュールを選択する学習ができていると言える．

7　おわりに

本研究では複数の応答生成モジュールを組み合わせて雑談対話を行うシステムにおいて，その選択のために強化学習を用いることで適切な対話戦略を獲得することを提案した．実際に対話を重ねることで，適切に応答生成モジュールを選択するような対話戦略を獲得できることを示した．

参考文献

[1] E. Levin, R. Pieraccini, and W. Eckert. Using markov decision process for learning dialogue strategies. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 201–204. IEEE, 1998.

[2] T. Meguro, R. Higashinaka, Y. Minami, and K. Dohsaka. Controlling listening-oriented dialogue using partially observable markov decision processes. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 761–769. Association for Computational Linguistics, 2010.

[3] S. Young, M. Gasic, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu. The hidden information state model: a practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, Vol. 24, No. 2, pp. 150–174, 2010.

[4] 横山祥恵, 山本大介, 古賀敏之, 小林優佳, 土井美和子. 高齢者向け対話インタフェースの開発: 概念辞書を用いた話題展開法. 情報処理学会全国大会講演論文集, Vol. 71, No. 4, 2009.

[5] 翠輝久, 河原達也, 正司哲朗, 美濃導彦. 質問応答・情報推薦機能を備えた音声による情報案内システム. 情報処理学会論文誌, Vol. 48, No. 12, pp. 3602–3611, 2007.