音声発話系列からのユーザの意図の理解

User Intention Understanding from Utterance Sequences

吉野幸一郎

Abstract

音声言語理解を用いるアプリケーションにおいて，複数のスロットを持つような複雑なユーザ意図の理解やタスクの実現には，音声発話の系列を考慮することが不可欠である．こうした系列の考慮は，音声言語理解を用いたアプリケーションの可能性が示唆された頃から検討がされており，様々なアプローチがとられてきた．本稿では，こうしたアプローチを概観するとともに，分野として今後どのような発展が期待されるかを解説する．

キーワード：対話状態推定，対話履歴

1. ま え が き

システムとユーザが自然なインタラクションを行う場合，複数のやり取りが発生するが，このような場合は個々のユーザ発話のみからユーザの意図全体を認識することが難しい．こうした場合，ユーザの意図を正しく認識するためには，やり取りの履歴，つまり音声発話の系列全体からユーザの意図を推定する必要がある．本小特集第1章で述べられているように，ユーザの意図はフレームで表現される場合が多いが，このフレームで扱うスロット数が多いドメインでは，ユーザが自身の意図を1発話で全て伝えることは難しく，複数の発話にまたがって一つの意図が伝えられるためである．このような意図の認識を行う場合，単に各発話から得られるユーザ意図を個別に認識した上で結合するだけでは不十分で，過去に入力された内容を考慮して現在の意図理解結果を修正する，応答フィルタのようなモデルを用いる必要がある．対話システム研究の分野においては，こうした発話系列を考慮した意図理解結果を対話状態と呼ぶ．本来こうした枠組みの利用は対話に限らないものであるが，本稿では先行研究に倣い，対話状態の語を用いる．また，このような意図理解結果の修正を行う問題を，特に対話状態推定と呼ぶ(1)．

対話状態推定のように発話系列を考慮する場合，状態をどのような粒度で持つのかによって，考慮しなければならない系列の範囲が異なる場合がある．ここでは，フレームのようなslot-valueの組合せでユーザの意図を保持する場合を例に取る．フレームはこれまでの音声言語理解の枠組みで多く用いられてきたが，こうした表現が有効なのは達成すべき目標が明確であるようなアプリケーション，例えばレストラン案内やカーナビゲーションシステム，などの特定のタスクを達成する対話システムが多い(2)～(4)．この場合，フレームの種類や話題に相当するスロットは長期にわたって考慮し続ける必要がある．例えば，ユーザがレストランについて検索したいと思った場合，レストラン案内について記述したフレームは，レストラン案内が継続する限り用いなければならない．それに対し，より短期で入れ替わるようなスロット値については，直近からのユーザ発話からの情報をより重点的に考慮しなければならない．

このように，発話間での影響度合いはタスクやドメインによって様々であるが，以前の発話の影響を，システムが必要に応じて選択可能であることが望まれる．こうした意図理解における系列の考慮は，特に対話システムの分野においては古くから論じられており(5)，近年の機械学習を用いた統計的予測によって急速に精度が向上した(6)～(8)．また，近年の深層学習研究の進展から，機械学習を用いることができる幅が急速に広がっており，その発展に応じて部分問題の定義や範囲も変化している．

吉野幸一郎　奈良先端科学技術大学院大学先端科学技術研究科
E-mail koichiro@is.naist.jp
Koichiro YOSHINO, Nonmember (Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma-shi, 630-0192 Japan).
電子情報通信学会誌 Vol.101 No.9 pp.896-901 2018年9月

例えば，各発話に対する言語理解と，個々の理解結果を統合する対話状態推定は，別個のモデルをパイプライン的に処理するのではなく，ある時点までのユーザ発話系列から直接現在の対話状態を推定するような問題へと変化している．本稿ではこれらの問題定義や用いられるアプローチについて概観するとともに，こうした分野で今後期待される進展について議論する．続く 2. では，まず 2.1 で古典的なフィルタモデルによる意図理解の更新について述べ，続いて 2.2 で深層学習的アプローチについて述べる．また，発話から特徴量ベクトルを構築しニューラルネットワークに入力する手法について 2.3 で解説する．また，これらにおける研究の現状とこれからの展望について，それぞれ 3. と 4. で述べる．

2. 意図理解における履歴の考慮

一般に音声系列からの意図理解を行う場合，図 1 に示すように，1 発話に対する音声認識，言語理解の処理をパイプライン的に行った上で，言語理解の出力結果に対して履歴の考慮を行う．例では，ユーザの「1. 東西線で難波まで行きたい」「2. 東西線で馬場まで行きたい」という複数の音声認識結果候補からスロットを抽出した上で，対話履歴にある「乗る駅が早稲田である」という情報を考慮して最終的な意図理解の結果を出力している．こうした入力履歴の考慮において一番単純な手法は，抽出したスロットを毎回理解結果の上書きに用いることであり，対話システム研究の初期においてはそのような実装が行われる場合も多かった．

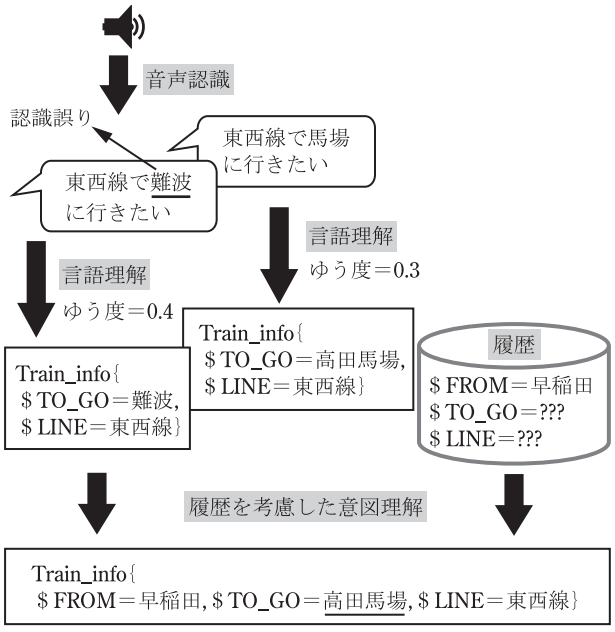

図 1　音声言語理解における発話系列の考慮

2.1　フィルタモデルによる意図理解の更新

しかしそのような単純な方法では，履歴を用いることが適当な場合と不適当な場合の判断ができない．また，音声認識や言語理解の誤りに由来する間違った入力がなされた場合への対処も問題となる．近年の音声認識や言語理解は確率モデルに基づいており，各発話の理解結果に対するゆう度を出力することができるため，このゆう度やスロット同士の共起確率などを考慮して理解結果を修正することが必要である．

そこで，応答フィルタに代表されるようなフィルタモデルの枠組みを導入することで，この問題の解決が試みられている．発話系列上の t 番目の発話に対して言語理解モデルが出力した結果を o^t とし，これに対して履歴を考慮した上で得られるはずの真の意図が s^t であるとする．s^t は n 個の理解結果候補が存在し，このうち i 番目のものを s_i^t とする．このとき，各状態 s_i^t に対する確率は，

$$P(s_i^t|o^1, o^2, \cdots, o^t) \propto P(o^t|s_i^t)\sum_j P(s_i^t|s_j^{t-1})P(s_j^{t-1}|o^{1:t-1}) \quad (1)$$

として与えることができる(注1)．一般に $P(o^t|s^t)$ は観測確率と呼ばれ，音声認識及び言語理解のゆう度で与える（$P(o^t|s_i^t) \approx P(o^t)$）．また $P(s_i^t|s_j^{t-1})$ は状態遷移確率と呼ばれ，コーパスなどから計算する．このように，誤差のある観測値から真の状態を予測する枠組みは，正にフィルタにあたる．

こうした枠組みを用いることで，仮に音声認識や意図理解の結果が誤っていたとしても，履歴からその誤りを是正することができる．例えば，共起しづらい二つのスロット（早稲田駅から難波駅など）が埋まり，かつ観測確率（ゆう度）も低いような場合に，これを是正する（難波を馬場に修正する）などの効果が期待できる．

このように，発話系列を考慮した意図理解を，対話システムの分野では特に対話状態推定と呼ぶ．対話状態推定は対話システム研究の重要な課題として位置付けられており，特にタスク達成対話において様々なデータでのチャレンジタスクが開催されている(1), (11)～(14)．これらのチャレンジタスクでは，初期は各時刻の意図理解結果とそのゆう度を入力としていたが，第 4 回以降は発話の単語系列を入力として用いている(注2)．そうした意味で，意図理解（言語理解）と対話状態推定の問題設定の境界は曖昧となりつつある．

（注 1）　意図理解の後段処理である対話制御に強化学習を用いる場合，時刻 t におけるシステムの行動 a^t の系列も含めてこの更新関数を定義する(9), (10)．今回はユーザの発話のみを考慮する場合を扱うが，本稿で解説するモデルはいずれもシステムの行動（発話）を用いる拡張が存在する．

（注 2）　ただしこれらのチャレンジでは発話の書き起こしが与えられている．実際に音声認識結果を用いる場合，各認識仮説をゆう度で重み付けした n-best 候補を用いることが多い．

2.2 リカレントニューラルネットワーク（RNN）を用いた意図理解の更新

ニューラルネットワークを用いた技術の進展は，ユーザ発話の意図理解にも影響を及ぼしている．特に音声発話系列からの意図推定においては，時系列的な変化を捉えるため，再帰形の構造を用いることが一般的に行われている（RNN：Recurrent Neural Networks）．この場合，言語理解とフィルタモデルは分離せず，一体的に入力単語系列から発話系列を考慮した意図の推定を行うことができ，精度の向上が期待できる．これは，履歴をどの場合にどの程度利用するかについても，意図理解の精度を最大化するように学習することが可能となるためである．

t 番目の発話から得られるベクトルを $\boldsymbol{x}^t$ とする（1 発話を一つのベクトルで表現する）とき，RNN の隠れ層 $\boldsymbol{h}^t$ は，図 2 のように表すことができる．具体的には，

$$\boldsymbol{h}^t=\tanh(W_{xh}\boldsymbol{x}^t+W_{hh}\boldsymbol{h}^{t-1}+c_h) \quad (2)$$

$$\tanh(z)=\frac{e^z-e^{-z}}{e^z+e^{-z}} \quad (3)$$

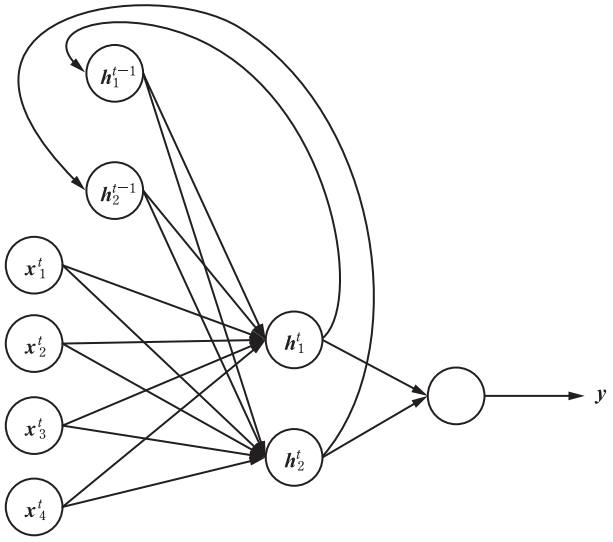

図 2　リカレントニューラルネットワーク（RNN）の構造

によって更新される．ここで W_{xh} は入力層 $\boldsymbol{x}$ と隠れ層 $\boldsymbol{h}$ の重み伝搬行列，W_{hh} は前の時刻 $t-1$ における隠れ層からの重み伝搬行列，c_h はバイアスである．式（2）ではベクトルの各次元を行列演算で求めた上で，各次元に対して式（3）で与えられる計算を行った結果を $\boldsymbol{h}^t$ とする．この更新された隠れ層を入力とし，

$$\boldsymbol{y}^t=\mathrm{softmax}(W_{hy}\boldsymbol{h}^t+c_y) \quad (4)$$

$$\mathrm{softmax}(z)=\frac{e^{z_i}}{\sum_k e^{z_k}} \quad (5)$$

によって現在の対話状態 $\boldsymbol{y}^t$ を予測する．ここで W_{hy} は隠れ層 $\boldsymbol{h}$ から出力層 $\boldsymbol{y}$ への重み伝搬行列であり，c_y はバイアスである．こちらも同様に，行列演算で得られるベクトルの各次元を式（5）の入力とする．

ここで式（1）で与えられた信念の更新式と式（2）で与えられたリカレントニューラルネットワークの式を比較すると，観測確率が重み伝搬行列 W_{xh} で，状態遷移確率が重み伝搬行列 W_{hh} で置き換えられているだけで，機能的には大きな差異がないことが分かる．

また，再帰構造を持つネットワークとして Long Short-term Memory Neural Network（LSTM）を用いる場合もあるが，ネットワークが持つ役割に大きな違いはない．LSTM は RNN と比較して時間経過による勾配消失の問題が抑えられており，より長期の系列を考慮したい場合に適しているとされる．

2.3 対話状態推定への発話の入力

RNN や LSTM のようなネットワークを用いて対話状態を推定する場合，ある時点 t の発話は複数の単語 $\boldsymbol{w}_{1:n}^t$ から成り，これらを一つの入力ベクトル $\boldsymbol{x}^t$ へと変換する必要がある．この変換には，幾つかの手法が考えられる．最も単純な手法は，単語列 $\boldsymbol{w}_{1:n}^t$ からベクトル $\boldsymbol{x}^t$ への変換を考慮せず，各単語の入力を図 3 のようにネットワークに直接組み込み，各発話の単語列を順番に RNN に入力する手法である．この手法では単語列末に文末記号を挿入し，文末記号が入力された時点で意図理解結果を出力するようにネットワークを構成する．しかしこうした手法は勾配消失の影響により，発話文長の影響を受

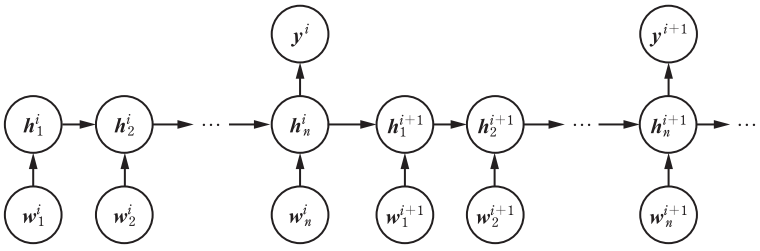

図 3　単純な単語列符号化に基づくモデル

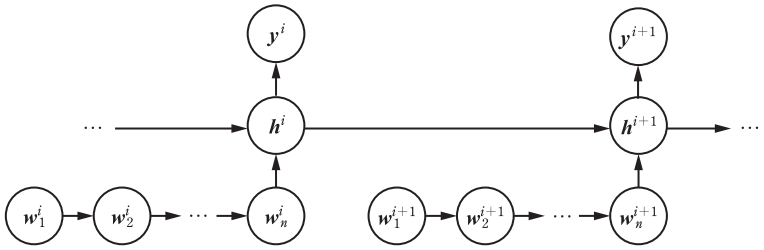

図 4　階層的な単語列符号化に基づくモデル

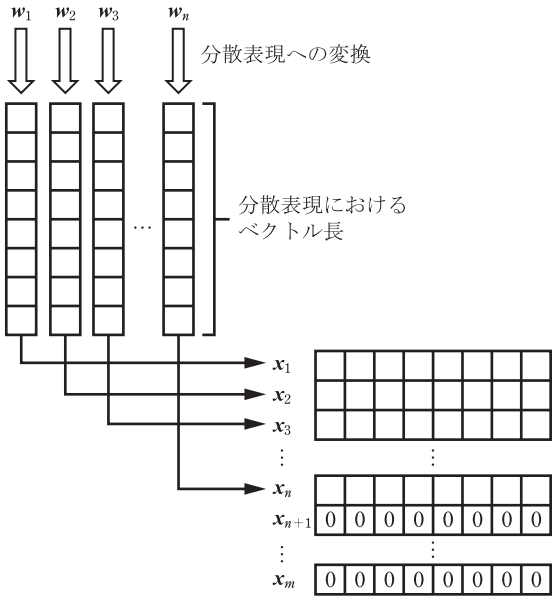

図 5　発話系列の行列への変換

けやすい．そこで図 4 のように，発話の各単語を順次入力していく層と，発話間の関係を考慮する層を明示的に分ける構成を取る場合もある．こうした構成は，特に Neural Conversation Model の研究では盛んに用いられている．

発話中の単語を順番に符号器に入力する単純なアプローチにおいては，単語をそのまま入力する場合と，word2vec[15] に代表される分散表現を用いて単語を固定長ベクトルに変換してから入力する場合が存在する．分散表現を用いる場合，単語から分散表現への変換は，Wikipedia などの大規模テキストデータから学習されたものを用いることが多い．

また，単語を順番に入力していく手法のほかに，Convolutional Neural Network（CNN）を用いて発話全体を符号化する手法がある．この手法では図 5 に示すように，固定長ベクトルに変換された各単語を行とし，各列に並べることで行列を構成し CNN の入力として用いる．この際列長 m は，学習データ中の最も長い発話に合わせて設定し，それより短い発話については 0-padding を行う．これらの手法のいずれがよいかについては議論があるが，筆者の経験としては分類問題においては CNN による圧縮を入力として用いるのが有効である．

3. 対話状態推定における研究の動向

系列を考慮した意図理解においては，特に対話状態推定のチャレンジタスクである Dialogue State Tracking Challenge（DSTC；Dialogue System Technology Challenge）[注3] や Air Travel Information Services（ATIS）[注4] タスクのオープンデータで様々な手法が提案されている．Dialogue State Tracking Challenge 2 においてリカレントニューラルネットワークを利用するアプローチが提案され[16]，これを LSTM に拡張したもの[17,18]や，各ターンにおいて CNN を利用するもの[19]が提案されている．

タスク達成対話における対話状態推定では，DSTC2 タスクにおいてスロット単位の F1 スコアは 94% を上回ることが報告されている[16]．つまり，既存のタスク達成対話においては，十分に統制のとれた（アノテーション一致率が高い）学習データが存在する場合，十分な精度を達成することができる．

一方で，十分に統制のとれた学習データを準備することが困難な場合，対話状態推定の精度が十分に達成されているとは言い難い．例えば DSTC5 においては，最高精度を達成したシステムでも各スロットの F 値は 45% 程度にとどまっている．これはデータそのものの統制（アノテーションの一致率）と，問題の難しさに対するデータの少なさの双方が問題と考えられる．このうち特にデータの少なさの問題に対しては，深層学習以前から転移学習や事前分布の仮定，ルールとの併用などの手法が取られており，そのような手法を適用していくことが求められる．

（注 3）　https://www.microsoft.com/en-us/research/event/dialog-state-tracking-challenge/
（注 4）　https://catalog.ldc.upenn.edu/LDC95S26

4. 系列を考慮した意図理解の今後の展望

音声発話系列に対する意図理解においては，機械学習，特に深層学習の適用により，用意された学習データの質，量が十分な場合は，人間のアノテーションと同等レベルの精度が実現されている．一方で，十分な精度が実現されているのは一部の質，量ともに十分な学習データセットが存在するドメインの意図理解に限られており，そのようなデータが存在しないドメインに対して既存の意図理解の枠組みをどう適用していくかということは大きな課題である．これは特に実製品に既存の手法を適用しようとする場合に問題となる．つまり，新規のタスク・ドメインをデザインした際に，学習データをどのように準備するかという課題が存在する．

ここまで述べた内容を愚直に実用システムで運用しようとする場合，まずデータ収集用のシステムを構築し，収集されたデータに対してアノテーションを行い，そこからモデル構築をするという手順が必要となる．ただ，スロットやフレームの追加や新規ドメイン運用のニーズに対してこれらの運用コストが見合うとは言い難い．この問題に対し，新規ドメインでのシステム運用をより簡単にするため，ルールと統計処理の双方を用いるハイブリッドと呼ばれるアプローチが存在する[20),21)]．こうした研究では，当初人手で記述したルールやドメイン知識を，シームレスに統計モデルへと移行する手法が検討されている．

また，新規ドメインに対して，マルチドメインシステムとドメイン間の転移学習により意図理解を構築しようというアプローチが存在する[22),23)]．こうした枠組みでは既存の意図理解に対応したデータをそのまま活用できるが，ドメインや意図理解の構造が大きく異なるタスクに対応することが難しい．これに対し，一般的な対話行為とドメイン知識を区別して意図理解を構築しようという取組みも存在する[24)〜26)]．

また，何をもって理解とするかということについての議論も必要である．現状の意図理解の枠組みでは，システムが可能な行動について記述し，そのタスクに応じて必要最小限の対話フレームを人手で作成している．例として，DSTC2 における対話データとその意図理解，システムの行動のアノテーションを図 6 に示す．この例では，レストラン案内に必要な対話フレームを定義し，このフレームに当てはめる形で意図理解としているが，こうしたフレームの定義にはコストが掛かる．こうしたフレームを，様々なドメインに適用可能な汎用的意味表現を用いて半自動的に構築することができれば，こうしたコストを抑えるとともに，人間が期待する真の理解に一歩近づくのではないかと考えられる．ここでいう汎用的意味表現とは，例えば意味役割（Semantic Role Label），述語項構造（Predicate Argument Structure），Abstracted Meaning Representation などである．

更に，当該ユーザ発話がどのような文脈で発せられたかを文脈として考慮することも重要である．本稿では主に対話履歴の扱いについて述べたが，考慮することで発話の意図に対する解釈が変化するものは，対話履歴以外にも存在する．例えば，会話の文脈に応じて同じ発話がポジティブ，ネガティブ両面で解釈できるような場合，どちらの意味なのかは対話履歴のみでは判断がつかない場合がある．また，近年盛んに利用されているスマートフォンでのパーソナルアシスタントなどでは，機器の位置情報や操作履歴なども文脈として用いることが可能であり，これらを実際に用いている場合も多い．これには，例えば天気を尋ねると現在地周辺の気象情報を答え

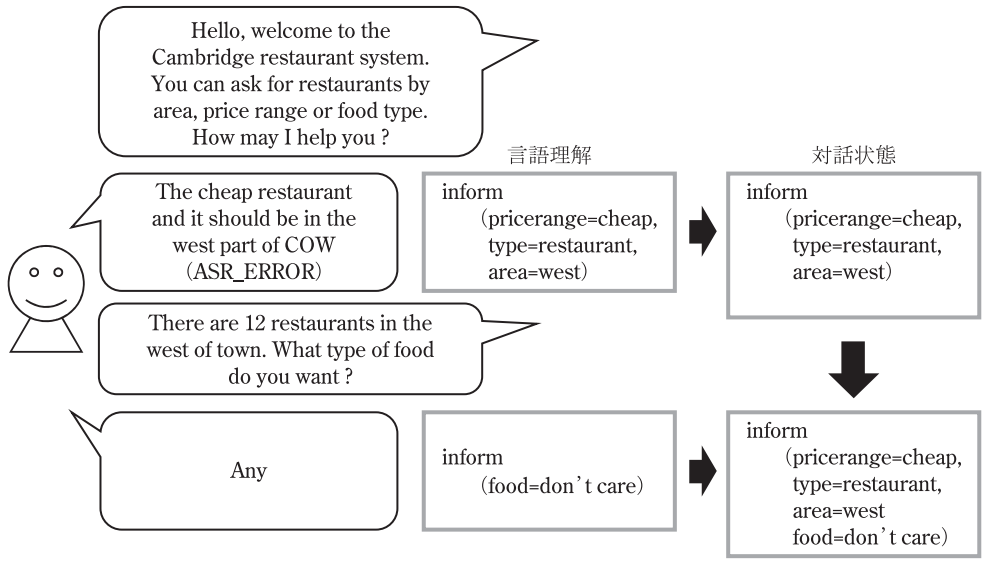

図 6　DSTC2 タスクにおける対話例と言語理解結果，対話状態推定結果

るなどの行為が相当する．こうした広義の文脈の考慮においては，これまで議論されてきた意図理解の枠組みをもう一度議論し直し，どこまで文脈として取り扱うことを期待されているかを再確認することが必要であると考えられる．

文　　献

（1） J. Williams, A. Raux, D. Ramachandran, and A. Black, “The dialog state tracking challenge,” Proc. SIGDIAL 2013 Conference, pp. 404-413, 2013.

（2） D.A. Dahl, M. Bates, M. Brown, W. Fisher, K. Hunicke-Smith, D. Pallett, C. Pao, A. Rudnicky, and E. Shriberg, “Expanding the scope of the atis task : The atis-3 corpus,” Proc. workshop on Human Language Technology, pp. 43-48 1994.

（3） K. Komatani and T. Kawahara, “Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output,” Proc. 18th conference on Computational linguistics-Volume 1, pp. 467-473 2000.

（4） S. Young, M. Gašić, 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,” Comput. Speech Lang., vol. 24, no. 2, pp. 150-174, 2010.

（5） S. Young and C. Proctor, “The design and implementation of dialogue control in voice operated database inquiry systems,” Comput. Speech Lang., vol. 3, no. 4, pp. 329-353, 1989.

（6） J.D. Williams and S. Young, “Partially observable markov decision processes for spoken dialog systems,” Comput. Speech Lang., vol. 21, no. 2, pp. 393-422, 2007.

（7） G. Mesnil, X. He, L. Deng, and Y. Bengio, “Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding,” Interspeech, pp. 3771-3775, 2013.

（8） K. Yao, G. Zweig, M. -Y. Hwang, Y. Shi, and D. Yu, “Recurrent neural networks for language understanding,” Interspeech, pp. 2524-2528, 2013.

（9） S. Young, M. Gašić, B. Thomson, and J.D. Williams, “Pomdp-based statistical spoken dialog systems : A review,” Proc. IEEE, vol. 101, no. 5, pp. 1160-1179, 2013.

（10） K. Yoshino, “Spoken dialogue system for information navigation based on statistical learning of semantic and dialogue,” PhD thesis, Graduate School of Informatics, Kyoto University, 2014.

（11） M. Henderson, B. Thomson, and J.D. Williams, “The second dialog state tracking challenge,” Proc. 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue（SIGDIAL）, pp. 263-272, 2014.

（12） M. Henderson, B. Thomson, and J.D. Williams, “The third dialog state tracking challenge,” Spoken Language Technology Workshop（SLT）, 2014 IEEE, pp. 324-329, 2014.

（13） S. Kim, L.F. D’ Haro, R.E. Banchs, J.D. Williams, and M. Henderson, “The fourth dialog state tracking challenge,” in Dialogues with Social Robots, pp. 435-449, Springer, 2017.

（14） S. Kim, L.F. D’Haro, R.E. Banchs, J.D. Williams, M. Henderson, and K. Yoshino, “The fifth dialog state tracking challenge,” Spoken Language Technology Workshop（SLT）, 2016 IEEE, pp. 511-517, 2016.

（15） T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” Advances in neural information processing systems, pp. 3111-3119, 2013.

（16） M. Henderson, B. Thomson, and S. Young, “Word-based dialog state tracking with recurrent neural networks,” Proc. 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue（SIGDIAL）, pp. 292-299, 2014.

（17） L. Zilka and F. Jurcicek, “Incremental lstm-based dialog state tracker,” 2015 IEEE Workshop on Automatic Speech Recognition and Understanding（ASRU）, pp. 757-762, 2015.

（18） K. Yoshino, T. Hiraoka, G. Neubig, and S. Nakamura, “Dialogue state tracking using long short term memory neural networks,” Proc. International Workshop on Spoken Dialog System Technology, pp. 1-8, 2016.

（19） H. Shi, T. Ushio, M. Endo, K. Yamagami, and N. Horii, “A multichannel convolutional neural network for cross-language dialog state tracking,” Spoken Language Technology Workshop（SLT）, 2016 IEEE, pp. 559-564, 2016.

（20） J.D. Williams, “The best of both worlds : Unifying conventional dialog systems and pomdps,” Ninth Annual Conference of the International Speech Communication Association, pp. 1173-1176, 2008.

（21） K. Yoshino, S. Watanabe, J. Le Roux, and J.R. Hershey, “Statistical dialogue management using intention dependency graph,” Proc. Sixth International Joint Conference on Natural Language Processing, pp. 962-966, 2013.

（22） M. Gašić, N. Mrkšić, P.-h. Su, D. Vandyke, T.-H. Wen, and S. Young, “Policy committee for adaptation in multi-domain spoken dialogue systems,” Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 806-812, 2015.

（23） A. Papangelis and Y. Stylianou, “Single-model multi-domain dialogue management with deep learning,” Proc. International Workshop on Spoken Dialogue Systems, pp. 1-6, 2017.

（24） T. Shibata, Y. Egashira, and S. Kurohashi, “Chat-like conversational system based on selection of reply generating module with reinforcement learning,” in Situated Dialog in Speech-Based Human-Computer Interaction, pp. 63-69, Springer, 2016.

（25） K. Yoshino, Y. Suzuki, and S. Nakamura, “Information navigation system with discovering user interests,” Proc. Annual SIGdial Meeting on Discourse and Dialogue, pp. 356-359, Aug. 2017.

（26） S. Keizer and V. Rieser, “Towards learning transferable conversational skills using multi-dimensional dialogue modelling,” Proc. SEMDIAL, p. 158, 2017.

（平成 30 年 4 月 18 日受付　平成 30 年 5 月 10 日最終受付）

吉野（よしの）　幸一郎（こういちろう）

2009 慶大・環境情報卒．2011 京大大学院情報学研究科修士課程了．2014 同博士後期課程了．同年日本学術振興会特別研究員（PD）．2015 奈良先端大・情報科学研究科・特任助教．2016 から同助教．同年から科学技術振興機構さきがけ研究員（兼任）．京大博士（情報学）．音声言語処理及び自然言語処理，特に音声対話システムに関する研究に従事．2013 年度人工知能学会研究会優秀賞受賞．IEEE，SIGDIAL，ACL，ISCA，情報処理学会，言語処理学会各会員．