音声対話システムの進化と淘汰

A Brief History of Spoken Dialogue Systems

河原 達也 [1]

Tatsuya Kawahara[1]

[1] 京都大学 学術情報メディアセンター

[1] Academic Center for Computing and Media Studies, Kyoto University

Abstract: Research and development of (spoken) dialogue systems have been conducted for almost 50 years, and significant progress has been made to realize a number of practical applications. This article first gives a brief overview on the history of spoken dialogue systems, and then point out the key factors for their evolution and selection. Finally, recent technical trend is also reviewed.

1. はじめに

最近、音声対話システムがブームになっている。米国では 2011 年秋に、Apple の iPhone に搭載された Siri が大きなセンセーションを引き起こし、日本では 2012 年春に、Siri だけでなく、NTT ドコモのスマートフォンで「しゃべってコンシェル」のサービスが開始され、音声対話システムの到達点として広く一般に認識されている。

長らく当該分野の研究開発に従事してきた者（著者を含む）にとっては、「ついにここまで来たか」という感慨とともに、「このブームは定着するのだろうか？」という一抹の不安も感じているのではないだろうか[Pieraccini 12]。音声対話システムは、本格的に研究開発されるようになってから少なくとも 20 年（これは著者の研究キャリアと符合する）の間に、進化と淘汰を繰り返してきた。本稿では、この歴史を振り返るとともに、音声対話システムが定着していくための条件と研究開発の展開について展望する。

なお音声対話システムとは、ユーザの発話した言葉を理解し、適切に応答するシステムであり[河原 06, 河原 04]、単純な機器操作（コマンド&コントロール）や、発話した言葉をそのまま検索エンジンに投げて結果を表示する音声検索は除外する。

2. 音声対話システムの歴史と系譜

これまでの代表的な音声対話システムの系譜を図 1 に示す。わかりやすさを優先して、システムの名称とプロジェクト名が混在している。他に重要なものも多数あるが、紙面の都合上すべて挙げられない点了解されたい。

2.1. コンピュータによる原型

対話システムの源流は 1960 年代の ELIZA[Weizenbaum 66]と SHRDLU[Winograd 72]に遡ることができる。ELIZA が単語の表層的なマッチングに基づいて応答を返すのに対して、SHRDLU は積み木の世界に限定しながらも深い概念理解を行った上で対話を行うものであった。後者の人工知能的なアプローチは Allen らの TRAINS などでより現実的な設定を対象としたシステムに発展している。ただし、これらは基本的に音声をインタフェースとして用いていない（TRAINS の後半では音声対応になっている[Sikorski 96]）。

これに対して音声認識・合成技術の発展を受けて、1990 年頃から本格的な音声対話システムが構築されるようになった。その先駆けは、MIT の VOYAGER[Zue 91]である。これは、ケンブジッジの街の案内を行うもので、タスクドメインの観点からも現在スマートフォンで行われているサービスに近い。ほぼ実時間で応答を行うシステムは音声研究コミュニティに大きなインパクトを与え、米国では DARPA 主導で 1990 年代前半に ATIS プロジェクト[Price 90]が行われた。フライトの情報案内タスクを対象として、米国の主要な研究機関が参画し、データ収集からシステム評価まで協調と競争の原理に基づいて行われた。現在最先端の統計的な言語理解や対話制御のモデルの源流は ATIS の終盤で考案されている。

日本でも 1990 年代前半に、音声対話システムの研究が活発に行われた。その先駆けは東芝の TOSBURG[坪井 91]である。ハンバーガショップのエージェントという設定は今では考えられないもの

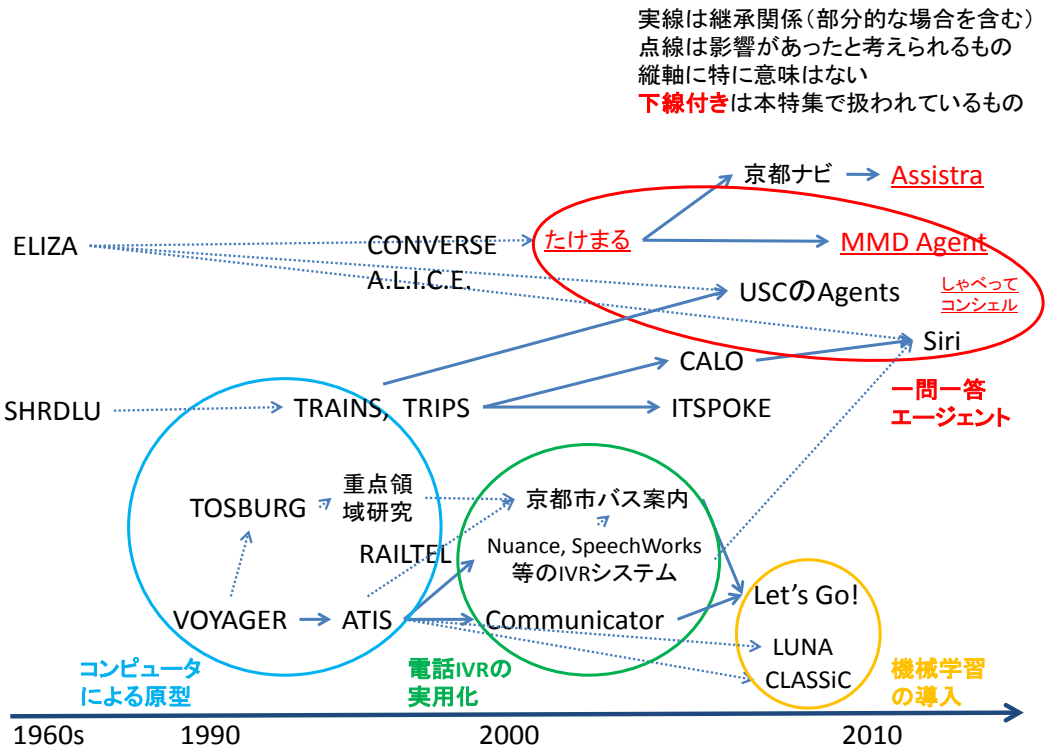

図 1　代表的な音声対話システムの系譜

であるが、自由発話に対してキーワードスポッティングを行い、実時間で動作するシステムは画期的であった。その後、科研費重点領域研究「音声対話」プロジェクト[堂下 98]が行われ、多数の大学の研究者が参画した。大学だけでなく、企業等の研究所も含めて、ほとんどすべての主要研究機関で音声対話システムが構築された。今から振り返っても、我が国で最も活発に音声対話研究が行われた時期であった。

2.2. 電話 IVR の実用化

その一方で、ATIS に従事していた研究機関の一部研究者がスピンオフして、Nuance 社や SpeechWorks 社を設立し、2000 年頃に相次いで株式上場を果たすと、新たな音声対話システムのブームを迎えることになった。これらは、電話（音声通話）のサービスに特化し、人手で記述された定型的な文法と対話フローに基づいて対話を行うものであった。すなわち、電話 IVR（自動音声応答）システムの入力をテンキーから音声発話に置き換えたにすぎないが、コールセンターで大規模に導入され、商用的に成功した。

このビジネスは日本でも展開されたが、結局のところ定着しなかった。その理由は、我が国では、人間による丁重な顧客対応を重視する傾向があることに加えて、早くから携帯電話によるネットアクセスが普及していたことなどが考えられる。著者らは、オムロン社の協力を得て、京都市バスの運行情報案内を行うシステムを開発し[Komatani 03]、京都市交通局との折衝を経て試験運用の形で 2003 年から一般公開した。ちなみにこれは、著者が SpeechWorks 社を訪問し、「音声対話システムは差し迫ったニーズに対応すべき」という議論をした中で着想したものである。約 4 年間運用したが、後期の主な利用者は視覚障害者に限定された。これもネットアクセス、特に二次元バーコードとの競合が最大の理由であった。

2.3. 機械学習の導入

米国における研究プロジェクトしては、2000 年代初頭に DARPA Communicator[Pellom 00]という ATIS をマルチドメインに発展させたものが行われたが、それ以降で最も顕著なものは、CMU で開発された Let's Go!システム[Raux 05]と思われる。これは、著者の研究室を経て CMU に移った Raux が、バスの時

刻表案内を行うシステムをピッツバーグ市域に展開したもので、人間のオペレータがいない夜間に正式なサービスとして運用されている。技術的には、Communicator で開発されたプラットフォームをベースにしている。大規模な対話データを収集でき、しかもオンラインのシステムを入れ替えることもできる利点を活かして、最近では一定の条件で外部の研究者に開放してコンペ型の研究開発を行っている。

欧州では、LUNA や CLASSiC などの音声対話システムに関する基礎研究プロジェクトが継続的に行われている。識別モデルに基づく言語理解[Raymond 07]や、POMDP に基づく対話制御[Young 06]など、統計的機械学習によるアプローチの先進的な研究が行われている。

2.4. 一問一答エージェント

上記に対して、深い言語理解や対話制御を行わない ELIZA 型の音声対話システムが 2000 年代になって再び脚光を浴びるようになった。これは、ベクトル空間モデル（詳細は 4.2 節参照）に基づく情報検索の影響もあると考えられる。我が国におけるその先駆けは、奈良先端科学技術大学院大学で開発された「たけまるくん」である[西村 04, 西村 12]。公民館・駅やイベントで公開されるエージェントとして、使われ続けている。このプラットフォームは、京都大学の「京都ナビ」[翠 07]や名古屋工業大学で開発されている MMD Agent[大浦 12]でも一部使われている。「京都ナビ」では、入力を応答テンプレートとマッチングするだけでなく、京都の観光地に関する文書集合を用意して、情報検索や質問応答に基づいた対話を行っている。「京都ナビ」の開発を主導した翠は、その後 NICT で京都の観光情報をスマートフォンで案内する Assistra[翠 12]の開発に従事している。NTT ドコモの「しゃべってコンシェル」[辻野 12]を含めて、近年注目を浴びている音声対話システムの大半がこの一問一答エージェントの範疇に属する。

米国でも、Allen の系譜をくむ対話システムの研究は、CALO プロジェクト[南 12]や USC の Traum らの会話エージェント[Leuski 10]に受け継がれているが、徐々に ELIZA 型の応答生成を取り入れるようになっている。その最たるものが、CALO プロジェクトからスピンオフした Siri である。Siri はスマートフォンの音声対話システムで、音声インタフェースに関しては Nuance のものを使用していると言われている。図 1 からもわかるように、Siri は音声対話システムの代表的な 3 つの系譜を受け継いでいるといえる。

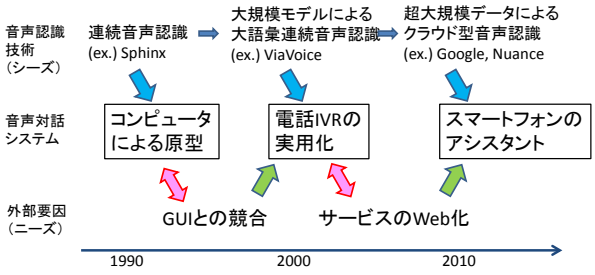

図 2　音声対話システムの進化と淘汰の要因

3. 音声対話システムの進化と淘汰の要因（図 2 参照）

図 1 で示した音声対話システムの歴史を俯瞰して興味深い点は、ブームが 1990 年頃（VOYAGER 等）、2000 年頃（Nuance 等）、2010 年頃（Siri 等）と約 10 年おきに生じていることである。

3.1. 音声認識・合成の進歩

この最大の要因としては、各々の数年前に音声認識システムにおいてブレークスルーが起きていることが考えられる。すなわち、1990 年の少し前に CMU の Sphinx に代表される連続音声認識がリアルタイムで実現され、2000 年の少し前に IBM の ViaVoice に代表される大規模な統計モデルに基づく大語彙連続音声認識が実用化され、2010 年の少し前に Google の Voice Search に代表されるクラウド型の超大規模データに基づくシステムが実現され、各時点で認識精度の点でも大きく改善された。音声合成についても同様の進歩があったと思われる。例えば、10 年前の IVR システムでは、ナレータによる録音音声でないと満足のいくものでなかったが、現在はテキスト音声合成(TTS)システムでも十分な品質に達している。

3.2. 他のインタフェースとの競合

一方、これまでの音声対話システムの研究開発の歴史を振り返ると、実用化に至らなかった、あるいは実用化しても成功しなかった要因として、他のインタフェースとの競合が考えられる。

1990 年代前半のコンピュータ（パソコンではなく Unix ワークステーションが大半）で構築された音声対話システムは、技術のショーケースとして位置づ

けられても、実際にはマウスやキーボードを使った方が確実で容易であるタスクドメインが多かった。特に Windows 95 で GUI が本格的に導入されてからは、GUI への優位性を明確に打ち出せるものがほとんどなかった。

したがって、2000 年代初頭に隆盛した音声対話システムは、音声のみがコミュニケーションチャネルである電話 IVR システムを主なターゲットとすることで成功を収めた。しかし、情報案内や予約受付などの多くのサービスが音声電話から Web ベースに移行し、音声対話システムの活路を見出すのが困難になった。Web ブラウザを音声で操作する規格やシステムもいくつか作成されたが、一般に使われることはほとんどなかった。

ところが 2000 年代後半になって、スマートフォンが登場すると、状況の変化が生じた。Web ベースのシステムであるが、キーボードで入力するのが困難な状況が生じたのである。Siri や「しゃべってコンシェル」などはこのような時機を得たシステムであるが、今後利用がどのように推移していくか注目したい。

3.3. 音声対話システムが生き残る条件

以上の考察をふまえて、音声対話システムが使われ続ける（生き残る）条件を以下にまとめる。

(1) 他のモダリティに対する優位性
GUI やキーボードに比べて明らかに効率がよいことは不可欠であろう。ロボットやカーナビのように、GUI やキーボードが使えない状況は理想的である。

(2) 差し迫ったリアルタイムな要求
音声で要求するのは差し迫った場合が多い。例えば、現在向かっている目的地への行き方がわからず、地図や交通手段を知りたい状況は、差し迫った要求といえる。一方、長距離移動の手段やホテルを調べるのは、ゆっくり他のインタフェースで行える。

(3) 日常的に使われるキラーアプリの存在
人間はインタフェースにある程度以上慣れないと使い続けない。上記の要件を満たすアプリであっても、たまにしか必要ないものであれば、結局のところ使われない。Ford の SYNC が使われ続けているのは、車内から電話をかけるという日常的に使うアプリの存在が大きいとのことである。

(4) 音声認識精度
あらためて述べるまでもないが、期待通りに動作しないものは使われない。

(5) エージェント・ロボットのキャラクタ
話しかけたいと思わせる魅力的なキャラクタも鍵となる可能性がある。

我が国では 10 年以上前から、カーナビ・携帯電話・ロボットなど、(1)の条件を満たす状況で音声認識インタフェースが搭載されているが、それほど使われていないのは、(4)の音声認識精度の問題と、(2)(3)のアプリの問題ではなかろうか。

上記とは別にビジネスモデルの問題がある。結局のところ、ビジネスとして確立しなければ、試験運用はできても持続的なものにはならない。自動車やロボットなどの高価なものに搭載する場合は製品コストに積み上げることができるが、一般にインタフェースという性質上、エンドユーザから費用を徴収するのは容易でない。一方、電話 IVR システムなどを、まじめな顧客対応として導入してもらう場合は、非常に高い精度と品質が要求される。Siri や「しゃべってコンシェル」も無償のガジェット的サービスであることに留意されたい。

3.4. 対話管理の退化

対話管理については、研究レベルにおいては POMDP に代表されるように機械学習やシミュレーションを用いた洗練が行われているものの、近年の実用的なシステムにおいては、詳細かつ固定的なフローに基づくか、一問一答に徹するかになっている。最も典型的な点は音声認識誤りへの対応であり、ユーザ発話に対して逐一確認する保守的な戦略か、一切確認しない一問一答型かのいずれかになる。逐一確認する対話例を以下に示す。

（逐一確認を行う対話例）
S: 交通案内システムです。出発地を最寄りの駅名でおっしゃってください。
U: 奈良
S: 出発地は奈良ですか？「はい／いいえ」で答えてください。
U: はい
S: 目的地の駅名をおっしゃってください。
U: 京都
S: 目的地は京都ですか？「はい／いいえ」で答えてください。
U: はい
S: 奈良から京都へ行く電車は以下の通りです。….

このような対話を行うシステムは手間がかかり、ユーザにとって決して快適なものではない。しかし、予約受付などのトランザクションを行うシステムでは、入力間違いは一切許されないので、上記のような確実性が保証される設計にせざるを得ない。

京都市バス運行情報案内システムのように無償の試験サービスでは、基本的に自由な発話を受け付け、検索を行う前に一度だけ確認を行う設計にしていた。ただしこのような戦略も、項目が多くなると、誤りが含まれる可能性が増える反面、その項目を指定して訂正するのは容易でない。これは、Web でのフォーム入力と比較して、決定的に異なる点である。

ところが、最近スマートフォンで行われているこの種のサービスでは、基本的に確認を行っていない。地名が発話されれば、それを目的地とみなして、地図の表示や交通の案内を行う。出発地は現在地とみなして、スマートフォンの位置情報を取得することで認識している。もちろん「奈良から京都まで」といった発話がされると、それに応じた応答がされるが、この場合も確認は一切行わない。一問一答に徹することで、たとえ音声認識誤りがあったとしても再度あらためて発話してもらえばよいという考え方である。また、場所や時間に関する情報を取得してデフォルト値とすることにより、曖昧性を極力解消している。すなわち、コンテクスト情報を取得することで、できるだけ対話を回避しているのである。対話をできるだけステートレスにしようとするのは、対話制御の退化ともいえる。

4. 今後の展望

現状の音声対話システムを冷静に考察すると、情報検索はできていても、話し言葉で会話できているわけではないことがわかる。Siri や「しゃべってコンシェル」を利用するシーンを考えると、システムができることを事前に考えて（概念的制約）、単純な文を（言語的制約）、明瞭に話す（音響的制約）必要がある。要するに、機械を意識した発声をする必要があり、しかもほとんど一問一答しかできない。音声対話システムのめざす姿が、ホテルのコンシェルジェや観光地のガイドだとすると、相当のギャップがあるのは明白である。

4.1. 真の情報コンシェルジェに向けて

たとえ機械らしい対話でよいとしても、深く長い対話を行うシステムに向けた研究開発は重要であろう。そのために、ユーザの情報要求がキーワード検索のように明確でない場合でも、ユーザの意図や嗜好を対話を通じて明確化する枠組みが必要である。例えば、観光地を巡る場合やレストランを探す場合は、明確な目標を有しているとは限らない。そういったニーズに対応できるのが真のコンシェルジェであり、著者らは情報コンシェルジェというコンセプトで研究を行っている[河原 08]。これは、料理や工作などの支援を行う問題解決型対話や、健康や法律などの相談を行う対話にも通じると考えられる。

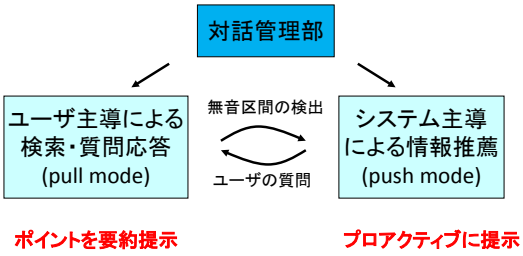

図 3 情報コンシェルジェにおける対話戦略

人間のコンシェルジェやガイドの最大の特徴は、ユーザの質問に答えるだけでなく、プロアクティブに情報を提示することである。音声対話システムを会話相手としてとらえた場合に、一問一答型のシステムでは、ユーザがきくことがなくなると会話が終了してしまう。実際に、「たけまるくん」や「京都ナビ」の運用例でも、大多数のユーザは数ターンで終了している。ただし、「京都ナビ」では、ユーザが発話しなくなると、システム側から積極的に情報推薦をする機構を用意しており、実際に 30 ターン以上対話した人も多数いる。この対話の枠組みを図 7 に示す[翠 07]。情報推薦をする際には、それまでの対話の履歴からユーザの嗜好・興味を推定し、それに沿った情報を提示することが重要である。

4.2. 情報提示型対話

現状の音声対話システムが、ユーザからの情報要求に対して応答するという受け身型のシステムであるのに対して、システム側からの情報提示を主とする対話の形態も考えられる。

例えば、観光地や博物館にあるオーディオガイドや、テレビやラジオのニュース番組のように、音声で情報を提示しているものをインタラクティブにすることが考えられる。その一環として、著者らはニュース記事の案内を行う音声対話システムを構築している[吉野 11]。これはニュースを提示しながら、ユーザの質問に応答することができ、しかも関連す

る話題も提示する機能がある。この対話例を以下に示す。

（ニュースナビの対話例）
S: 昨日のプロ野球の結果です。
　中日対阪神は、…
U: それお願いします
S: 中日は最終回に XX のホームランで逆転しました。
U: 阪神の誰が打たれたんですか
S: YYY が打たれました。
S: ちなみに○月△日には、YY は巨人・ZZ に逆転ホームランを打たれました。

e-Learning を音声対話によりインタラクティブにする試みも、Litman らの ITSPOKE において行われている[Litman 04]。音声対話を通じることにより、非言語情報からユーザの心的状態（理解しているか、集中しているか等）を推定することも可能になる。

5. おわりに

音声対話システムの歴史を振り返りながら、その進化と淘汰の要因について考察を行った。さらに最近実用化されている音声対話システムの特徴について述べた。

現時点が音声対話システムの研究開発において大きな到達点・分岐点になっていると考えられる。今後さらに高度化・発展し、世の中に浸透していくことを期待する。その際には、図 1 も改訂されることであろう。

謝辞

本稿の 2 章・図 1 の音声対話システムの歴史・系譜を編纂するに際して、荒木雅弘、駒谷和範、翠輝久の各氏から貴重な情報提供とコメントを頂きました。感謝申し上げます。

参考文献

[Pieraccini 12]　Pieraccini, R.: The Voice in the Machine. MIT Press (2012)
[河原 06]　河原達也, 荒木雅弘: 音声対話システム. オーム社 (2006)
[河原 04]　河原達也: 話し言葉による音声対話システム. 情報処理, Vol.45, No.10, pp.1027--1031, (2004)
[堂下 98]　堂下修司他（編）: 音声による人間と機械の対話. オーム社 (1998)
[Weizenbaum 66]　Weizenbaum, J.: ELIZA -- A computer program for the study of natural language communication between man and machine. Commun. ACM, Vol.9, No.1, pp.36--45 (1966)
[Winograd 72]　Winograd, T.: Understanding Natural Langauge. Academic Press (1972)
[Sikorski 96] Sikorski,T. and Allen, J.: A task-based evaluation of the TRAINS-95 dialogue system. In Proc. ECAI Workshop on Dialogue Processing in Spoken Language Systems (1996)
[Price 90] Price, P.J.: Evaluation of spoken language systems: the ATIS Domain. In Proc. DARPA Speech & Natural Language Workshop (1990)
[Zue 91] Zue, V., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J., and Seneff, S.: Integration of speech recognition and natural language processing in the MIT VOYAGER system. In Proc. IEEE-ICASSP, pp.713--716 (1991)
[坪井 91] 坪井宏之, 竹林洋一, 橋本秀樹: キーワードスポッティングに基づく連続音声理解. 信学技報, SP91-95 (1991)
[Bennacef 96] Bennacef, S., Devillers, L., Rosset, S., and Lame1, L.: Dialog in the RAILTEL telephone-based system. In Proc. ICSLP, pp.550--553 (1996)
[Komatani 03] Komatani, K., Adachi, F., Ueno, S., Kawahara, T., and Okuno, H.G.: Flexible spoken dialogue system based on user models and dynamic generation of VoiceXML scripts. In Proc. SIGdial, pp.87--96 (2003)
[Pellom 00] Pellom, B., Ward, W., and Pradhan, S.: The CU Communicator: an architecture for dialogue systems. In Proc. ICSLP, pp.723--726 (2000)
[Raux 05] Raux, A., Langner, B., Bohus, D., Black, A.W. and Eskenazi, M.: Let's Go Public! Taking a spoken dialog system to the real world. In Proc. InterSpeech, pp.885--888 (2005)
[西村 04] 西村竜一, 西原洋平, 鶴身玲典, 李晃伸, 猿渡洋, 鹿野清宏: 実環境研究プラットホームとしての音声情報案内システムの運用. 電子情報通信学会論文誌, Vol.J87-DII, No.3, pp.789--798 (2004)
[西村 12] 西村竜一他: 10 年間の長期運用を支えた音声情報案内システム「たけまるくん」の技術. 人工知能学会誌, Vol.28, No.1, 2013.
[翠 07] 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)
[翠 12] 翠輝久他: 音声対話による観光案内システムの開発と多言語化　—音声対話システム AssisTra の研究開発から得られた知見と課題—. 人工知能学会誌, Vol.28, No.1, 2013.
[大浦 12] 大浦圭一郎他: キャンパスの公共空間におけるユーザ参加型双方向音声案内デジタルサイネージシステム. 人工知能学会誌, Vol.28, No.1, 2013.
[辻野 12] 辻野孝輔他: 実サービスにおける音声認識と自然言語インタフェース技術. 人工知能学会誌, Vol.28, No.1, 2013.
[南 12] 南泰浩: 統計的手法による音声対話制御. 情報処理, Vol.53, No.10, pp.1088--1094 (2012)
[Leuski 10] Leuski, A. and Traum, D.: Practical language processing for virtual humans. In Proc. AAAI/IAAI (2010)
[河原 08] 河原達也, 川嶋宏彰, 平山高嗣, 松山隆司: 対話を通じてユーザの意図・興味を探り情報検索・提示する情報コンシェルジェ. 情報処理, Vol.49, No.8, pp.912--918 (2008)
[吉野 11] 吉野幸一郎, 森信介, 河原達也: 述語項の類似度に基づいてニュース記事の案内を行う音声対話システム. 人工知能学会研究会資料, SLUD-B102-08 (2011)
[Litman 04] Litman, D. and Silliman, S.: ITSPOKE: An Intelligent Tutoring Spoken Dialogue System. In Proc. HLT-NAACL Demo. (2004)