Vol.2016-MUS-111 No.4
2016/5/21

音声対話システムの概観とその機能

吉野 幸一郎[1,a)]

概要：近年、音声認識の性能向上からそのアプリケーションとして音声対話システムが注目されている。音声対話システムはその役割によっていくつかのパターンに類型化され、求められる役割によって一般的な構成も異なる。本発表では、どのような機能・役割が音声対話システムに期待されているか、またその想定のもとでどのような構成がなされてきたかについて概観する。また、音声認識結果を扱うシステムとして、どのような機能が求められるかについても述べる。

An Architecture and Functions of Spoken Dialogue Systems

Koichiro Yoshino[1,a)]

1. 音声対話システムの構成

音声対話システムは、ELIZA、SHRDLU などの人と会話する人工知能を起源とし、現在の対話システムもこれらに大きな影響を受けている。河原 [1] が示した音声対話システムの歴史に、近年増え続けている様々な対話システムのバリエーションを加えたものを図 1 に示す。現在存在する対話システムは、その機能により詳細は異なるものの、基本的には図 2 の構造を基本としている。

まず、音声認識モジュールではユーザの発話をテキストの形へ変換する。近年深層学習を用いた音声認識の改善により、この部分の精度が劇的に向上しており、音声対話システムのような音声言語処理アプリケーションへの期待感を醸成する背景となっている [2]。

次に、音声認識結果が言語理解モジュールへと送られる。言語理解モジュールでは音声認識結果をシステムが理解可能な形に落とし込む。例えば図 2 の例のように、「京都駅からバスに乗りたい」という発話を「乗車するバス停＝京都駅」という形で、対話システムが持つタスクの遂行に必要な情報へと変換する。この言語理解の研究が近年盛んに行われており、特に Dialogue State Tracking Challenge[3] という国際コンペティションで精度の向上が図られている。

[1] 奈良先端科学技術大学院大学 情報科学研究科
630-0192 生駒市高山町
[a)] koichiro@is.naist.jp

その次の対話制御モジュールでは、言語理解モジュールの解析結果とそれまでの対話履歴を用いて、最適な行動選択を行う。音声認識の性能は向上し続けているものの、その精度は 100% ではなく、音声対話システムには場面に応じた適切な聞き返しなどの行動選択を行うことが求められている。また、対話履歴を考慮することで、これまで話した内容の中からタスク遂行に必要なものをユーザに尋ねるなどの行為を行うことができる。こうした対話制御は古くは人手で記述したルールなどを用いて行っていた [4] が、近年強化学習を用いた統計的対話制御によって実現されることが増えている [5]。また、これらの統合によって対話システムをメンテナンスしやすくする、スケーラビリティを向上させるということも重要な研究課題である [6], [7]。

発話生成では、対話制御がどのような内容を話すかを確定した後で、どのようにその内容を言語化するかが課題となる。これまではテンプレートや用例によって、あらかじめ用意されたものから生成を行うような手法が検討されていたが、近年ニューラルネットワーク言語モデルを用いた言語生成の研究が盛んに行われている [8]。

音声合成においては、生成された言語を発話音声として合成する。この際、対話においてより自然な発声、イントネーションを実現することが今後の課題として挙げられる。

2. 音声対話システムの類型と機能

1 章で述べた対話システムの構成は一般的ではあるが、

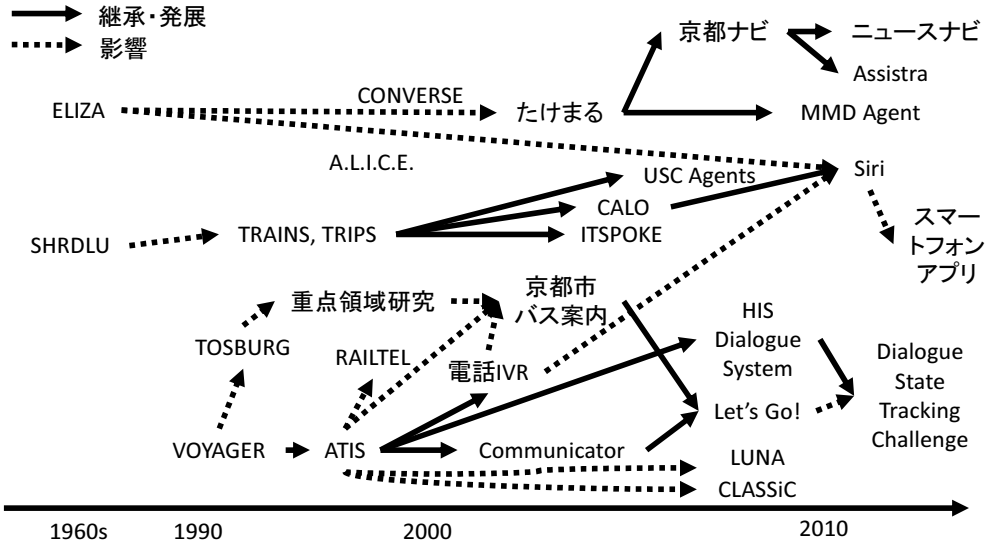

図 1　音声対話システムの歴史

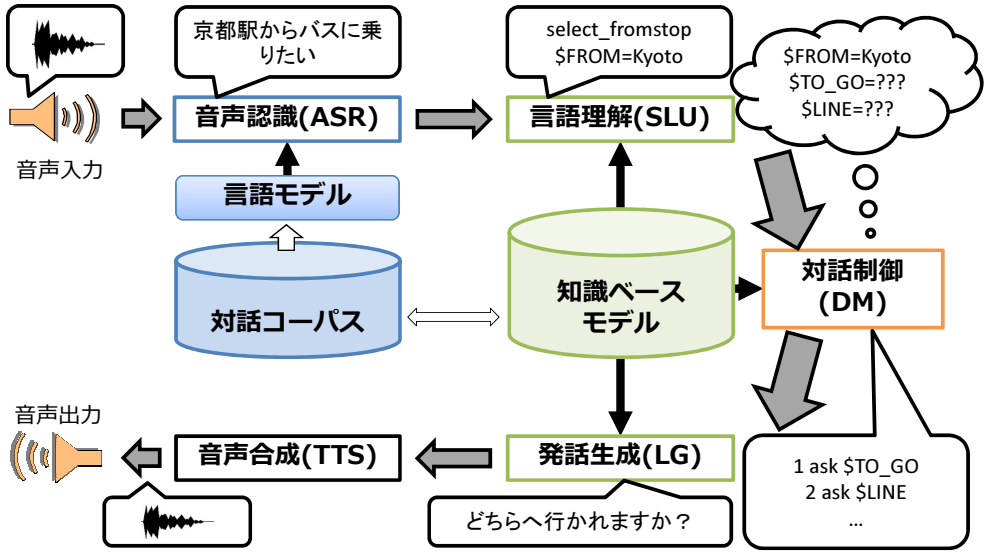

図 2　音声対話システムの構成

対話システムが置かれる状況の想定、求められる機能によってモジュールが省略される場合もある。例えば、雑談対話と呼ばれる特定のタスクゴールを指向しない対話においては、言語理解や対話制御を省略して end-to-end と呼ばれるアプローチをとることも多い [9]。end-to-end 型では多数のユーザ発話とシステム応答のペアを用意することにより、類似するユーザ発話に対応するシステム応答を応答文として用いたり、これを学習データとして生成を行ったりする。

これに対し、対話履歴を考慮し、特定のユーザゴールを達成することを目標として置いている対話システムをタスク指向型、ゴール指向型と呼ぶ。この他に、一度のユーザ発話、システム発話のやり取りで完結する一問一答型の対話システムも存在する。一般的な対話システムは、その目的に応じてこうした機能を複合することにより成り立っていることが多い。

3. おわりに

本稿では音声対話システムの一般的な構成と、それを用いる典型的な仕組みについて述べた。音声対話システムは様々なモジュールから成り立っており、その使われ方も様々であるが、目的に応じて適切に構成を作ることが必要となる。

また、これまでの対話システムの多くはユーザの音声発話区切りを 1 つの発話としてターンテイキングが成り立つことを想定している。しかし、実際のユーザ発話には言いよどみ、言い直しやさらに細かい単位での情報のやり取りが言語・非言語の双方で行われており [10]、これらを考慮することは今後の音声対話システム研究における重要な課題であるといえる。

参考文献

[1] 河原達也. 音声対話システムの進化と淘汰–歴史と最近の技術動向–. 人工知能学会誌, Vol. 28, No. 1, pp. 45–51, 2013.

[2] 久保陽太郎. 音声認識のための深層学習 (¡ 連載解説¿ deep learning (深層学習)[第 5 回]). 人工知能: 人工知能学会誌, Vol. 29, No. 1, pp. 62–71, 2014.

[3] Seokhwan Kim, Luis Fernando D'Haro, Rafael E. Banchs, Jason Williams, and Matthew Henderson. The Fourth Dialog State Tracking Challenge. In *Proceedings of the 7th International Workshop on Spoken Dialogue Systems (IWSDS)*, 2016.

[4] Deborah A Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. Expanding the scope of the atis task: The atis-3 corpus. In *Proceedings of the workshop on Human Language Technology*, pp. 43–48. Association for Computational Linguistics, 1994.

[5] Jason D Williams and Steve Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, Vol. 21, No. 2, pp. 393–422, 2007.

[6] Koichiro Yoshino, Shinji Watanabe, Jonathan Le Roux, and John R Hershey. Statistical dialogue management using intention dependency graph. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pp. 962–966, 2013.

[7] Kai Yu, Kai Sun, Lu Chen, and Su Zhu. Constrained markov bayesian polynomial for efficient dialogue state tracking. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, Vol. 23, No. 12, pp. 2177–2188, 2015.

[8] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1711–1721, 2015.

[9] Lasguido Nio, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. Recursive neural network paraphrase identification for example-based dialog retrieval. In *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, pp. 1–4. IEEE, 2014.

[10] Dan Bohus, Chit W Saw, and Eric Horvitz. Directions robot: in-the-wild experiences and lessons learned. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 637–644. International Foundation for Autonomous Agents and Multiagent Systems, 2014.