2004－SLP－53　(4)
2004／10／22

多言語音声対話システムアーキテクチャの比較検討

荒木雅弘
京都工芸繊維大学

あらまし 多言語音声対話システムのタスクとして観光案内を中心とした音声ポータルを設定した場合、従来研究以上の多言語への対応と保守性・拡張性の高さが実用化への重要な要件となる。本稿では音声対話システムの多言語化に関する技術を概観し、それぞれがカバーする領域を明らかにし、音声ポータル構築に適する手法を検討する。そして、これまで多言語音声対話システムで用いられてきた中間言語方式のアーキテクチャと、我々が提案する対話制御中心方式を比較し、提案方式の保守性・拡張性の高さを示す。

Comparison of architectures of multilingual spoken dialogue system

Masahiro Araki
Kyoto Institute of Technology

Abstract: In case the task domain of multilingual spoken dialogue system is fixed to voice portal which mainly deals with tourist information, it is necessary for practical systems to use maintainable and extensible method compared with previous systems. In this paper, we overview various methods for multilingual spoken dialogue systems and reveal the suitable problems for each methods. As a result, we compare intermediate language approaches to dialogue flow oriented approach, which is proposed in this paper, and show the high maintainability and extensibility of our proposed method.

1. はじめに

近年、音声翻訳の研究・開発が本格化し、多言語間での翻訳デモンストレーションも見られるようになった。外国語の習得や通訳なしに世界中の人々と会話ができる技術は究極の夢ではあるが、だれもが翻訳機を海外旅行に携帯し、どのような状況でも利用できるようになるにはまだしばらくの時間を要すると思われる。

一方、Web コンテンツへの音声アクセスをはじめとする音声対話システムは、その要素技術がある程度確立しており、目的を限定したものであれば、日々利用されているものも現存する(列車案内[1]や、バスの運行状況案内[2]など)。これらのシステムは便利ではあるが、同等の情報が携帯端末などの別手段で容易に入手できることもあって、利用が広がっているとは言えない状況である。

しかし海外からの旅行者が、自分の母語でこれらのシステムを利用したいという潜在的な要望は存在すると思われる。旅行者の全てがインターネットにアクセスできる端末を携帯しているとは限らないが、国際ローミングや空港での短期レンタルなどの手段で携帯電話を所持している旅行者は多い。

本稿では、旅行者に向けた音声ポータルを多言語で構築するための音声対話システムのア

ーキテクチャについて考察する。ここで多言語とは、最終的に 10 ヶ国語以上を想定する。さらに対象を主としてシステム主導の音声ポータルに絞る。この設定のもとで、多言語音声対話システムに適するアークテクチャを検討する。

以下、2 章では我々の考える多言語音声ポータルの設定を説明し、保守性・拡張性が最も重要な点であることを示す。3 章では多言語音声対話システムの先行事例を、特にアーキテクチャの観点からサーベイする。4 章では多言語音声対話システムを設計する際の論点を整理し、多言語音声ポータル構築に適した手法を検討する。5 章で我々が提案する対話制御中心方式のアーキテクチャを示し、これが多言語音声ポータルに適している点を説明する。6 章で本稿の議論をまとめ、今後の課題について述べる。

2. 多言語音声ポータルの設定

ここでは、我々が想定する多言語音声ポータルの構成について説明する。例えば、看板の文字が読めない、テレビニュースの内容がわからないような国に旅行したときには、交通手段・天候・宿泊・食事に関する情報などが音声対話で得られれば有益であろう。さらに観光ガイドも母語で受けられれば楽しみが増える。

このような設定を考えると、自動 call routing によって目的のタスクに応じた対話ができるようになるシステムが理想的である。しかし、自動 call routing には大量のデータが必要とされる[3]。これは多言語を対象とする我々の設定には適さないので、ここでは、最初にキーワードで情報ジャンルを選択し、次にジャンル毎に主としてシステム主導で対話を進める形式が妥当であると考えられる。混合主導対話は使い慣れているユーザには有益であるが、習熟度の低いユーザではタスク達成率が下がることや、時間はかかっても情報取得への動機が高いことなどを考えると、システム主導に限定されたユーザ主導(ヘルプなど)を加えたものが現実的な設定であると考えられる。

また、従来研究であまり重視されていなかった点として、保守性と拡張性を考慮する。ユーザの発話の多様性・想定外の遷移・タスクの追加など、音声対話システムでは運用後のシステムに対しても常に機能拡張を中心とした保守作業を行ってゆく必要がある。また、多言語システムとしては、最初は英語や中国語などのポピュラーな言語から実装し、ニーズを確認してからさらに多くの言語を追加するといった導入法が想定される。従って拡張性も重要になる。

3. 多言語音声対話システムの事例

3.1 MIT Voyager アーキテクチャ

Glass らは MIT で開発された音声認識・言語理解などの構成要素を組み合わせて、地域の案内を行う音声対話システム Voyager を開発し、その多言語化(日本語・イタリア語)を行った[4]。Voyager のアーキテクチャを図 1 に示す。

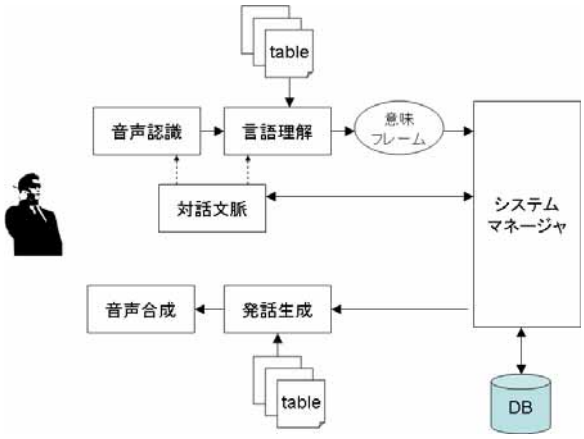

図 1 Voyager アーキテクチャ

言語理解部の枠組みとしては確率的な構文解析手法を採用しており、各言語に応じた構文規則と、学習用のコーパスが必要になる。また、発話生成は、意味表現からの文生成を行っており、語彙・文テンプレート・書き換え規則を各言語で用意すればよい。

基本的には言語依存の情報は中間言語によって吸収し、それより上位レベルの処理は言語独立で実装されている。中間言語は clause, topic, predicate の 3 項目を主要素としたもので、言語の違いが出てこないように設計されている。

3.2 Galaxy アーキテクチャ

Galaxy アーキテクチャ[5]は DARPA Communicator Program の参照アーキテクチャであり、図 2 に示すような Hub-and-Spoke 型の分散構成である。

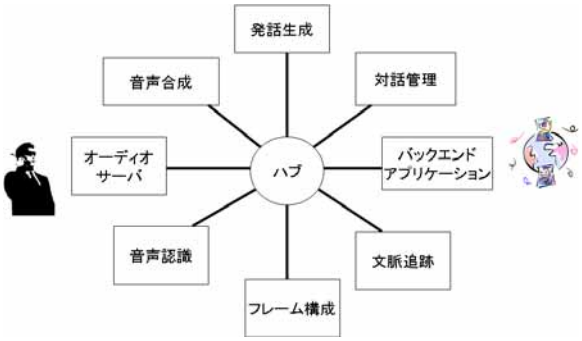

図 2 Galaxy アーキテクチャ

各分散モジュールは役割を与えられたサーバであり、ハブはサーバ間通信をサポートするルータの役目を果たす。サーバ間の通信はフレーム形式の言語で行われる。各サーバは plug&play の考え方に基づき、インタフェース部分を共通に実装すれば、任意のアルゴリズム・任意の言語で実装できる。

MIT では、この Galaxy アーキテクチャを用いて、天気情報ドメインにおいて日本語[6]、イタリア語、フランス語、ドイツ語などの音声対話システムが実装されている。このアーキテクチャでは、文脈追跡・対話管理・バックエンドアプリケーションのモジュールは言語独立に実装されている。

フレーム構成部は、音声認識結果(単語ラティス)を入力とし、言語独立な意味フレーム(E-form)を生成する。また、発話生成部は意味フレームを入力として応答文を生成する。

このような Hub-and-Spoke アーキテクチャはマルチモーダル対話システムなど複数のモジュールを高度に制御する際に有効になるもので、特に音声に特化してその多言語化の可能性を考える際は、制御情報ではなく言語情報がどのように流れるかを考慮する必要がある。言語情報の流れは Voyager システムとほぼ同じであり、中間言語が Voyager と比較してタスク寄りに設定されている点が異なる。

3.3 KIT アーキテクチャ

我々のグループでは、格フレーム変換による多言語音声対話システムのアーキテクチャを提案してきた[7]。これは、タスク依存の E-form を中間言語に設定すれば、各言語での同内容の表層表現を同じ中間言語に変換する規則の記述が難しくなることから、各言語での解析はそれぞれの言語での(タスク独立手法による)格構造抽出にとどめ、タスク依存の格フレーム変換処理と組み合わせることによって、見通しの良いアーキテクチャとなっている。

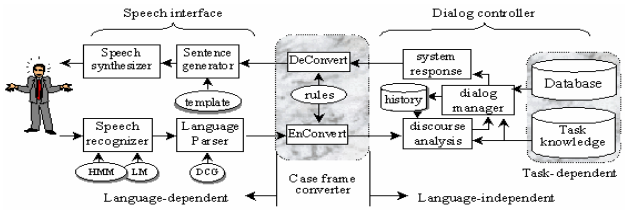

図 3 KIT アーキテクチャ

4. 設計上の選択肢の比較

3 章で取り上げた複数のシステムはいずれも言語理解(あるいは生成)部で何らかの中間言語を設定することによって、多言語システムを実現している。ここでは、これらを含め理論的に考えられる複数の方式を比較し、2 章で設定した我々のタスクである音声ポータルに適するアークテクチャを検討する。

4.1 音声認識・合成部の選択

3 章で概観したように、多くのシステムは対応する言語毎の音声認識・合成部を用いている。実装上の選択肢としてはこれらを共通化し、複数の言語を扱えるようにするという方式も考えられる。

認識部を共通化[8]すると、ユーザの第一発話によって言語識別が行えるので、利便性が向上する。また、記憶量に限界のある携帯端末上にシステムを実装する際には、メリットが大きい。しかし、認識結果は当然言語依存であり、その後の言語処理以降で言語の違いを吸収する必要がある。

合成部に関しては、モジュールそのものを共通化するのは上記の記憶量の点以外のメリットはあまりない。しかし、記述言語レベルで仮想的に共通化するという考え方はある。例えば、音声合成のための標準マークアップ言語である SSML [1] (Speech Synthesis Markup Language)では、句要素(<p>)や文要素(<s>)で発話したい言語を lang 属性で指定することで多言語出力に対応している。

4.2 発話理解・生成部の選択

アーキテクチャ上の選択肢としては明示的に中間言語を設定しない方法が考えられる。すなわち、発話理解・発話生成部を独立したモジュールとしては持たない方法である。

認識側では、部分的な認識結果を対話管理部が直接操作可能なオブジェクトにマッピングすることで、容易に複数の言語に対応することが可能である。しかし、この方式では、単語や句レベルで情報を抽出してしまうので、従来研究で扱っているようなある程度複雑な修飾関係を持つような文を扱うことは難しい。

一方、文生成モジュールを持たない場合は各言語に対応した応答文テンプレートを用意し、内容語に関しては音声合成器で対応可能なもの(数字、金額、日時など)、音素表記で対応可能なもの(固有名詞など)、オントロジーを利用するものなどを組み合わせて応答文を作成する。この方法では、高度な生成モジュールを利用したような自然な文の生成は難しい。

4.3 中間言語のレベルの選択

中間言語方式を採用する場合、設定する中間言語をタスク独立なものにするか、タスク依存のものにするかは設計上のポイントになる。

タスク独立なものにした場合は、機械翻訳に関する過去の研究を参考にし、複数の言語に共通する意味表現を抽出することになる。抽出規則は各言語の文法に依存したものになるが、属性文法などを利用し保守性の高い規則を得ることが可能である。しかし、タスク独立な意味表現から対話管理部で実際にアプリケーションを操作する言語(例えば SQL)に変換する必要があり、この部分の仕様が変更されると、前段階の中間言語抽出規則まで変更が及ぶ可能性がある。

一方、タスク依存の中間表現に直接変換する場合は、アプリケーション操作言語に直結しているという利点がある。しかし、他言語での対応する表層表現が同じ意味表現に変換されることを保証するように規則を保守することが

[1] http://www.w3.org/TR/speech-synthesis/

難しい。

5. 対話制御中心方式のアーキテクチャ

4.3 節で考察したように、他言語音声対話システムにおいて中間言語を設定した場合、その中間言語がタスク独立・依存に関わらず保守性の問題が生じる。一方、2 章で考察したような我々が設定する音声ポータルというタスクでは、タスクの追加・変更、言語の追加、対話フローのチューニングなど保守性・拡張性が最重要問題である。

ここでは、保守性・拡張性を重視したアーキテクチャとして対話制御中心方式を提案する。手法としては、アプリケーション部分以外は全て宣言的な知識記述を行うことによって多言語に対応させる。このことを可能にするために、標準記述言語と Web アプリケーションフレームワークおよびその多言語化手法を用いる。提案方式の概念を図 4 に示す。

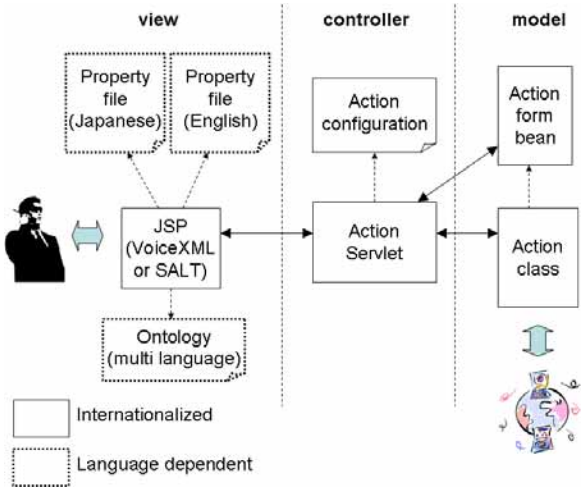

図 4 対話制御中心方式のアーキテクチャ

対話制御中心方式は MVC (Model- View-Controller) モデルに基づいており、タスクロジックとユーザインタフェース部が明確に分離されている。インタラクションによって得られる情報は図 4 の Action form bean に格納されるので、認識側はこの部分の言語独立性を実現するだけで多言語に対応できる。一方、生成側は Web アプリケーションの国際化対応手法を援用することができる。

この枠組みではユーザインタフェースは Web アプリケーションと相性の良い VoiceXML[2]あるいは SALT[3]を用いる。基本的にこれらの記述言語では Action form bean に格納する値をインタラクションによって直接取得することになるので、認識側での多言語の問題は文法記述に限定できる。認識は、単語および句レベルに限定し、文法規則にターゲット言語で利用するタグを記述することで多言語に対応する。従って中間言語に変換する手続きを記述する必要がないので、保守性を高めることができる。

また、生成側では Web アプリケーションの国際化手法を採用し、テンプレートを各言語のプロパティファイルに記述し、View にあたる JSP ファイルから参照する。また、内容語は表 1 に示すような処理で多言語化に対応する。

表 1 応答文生成における内容語の多言語化

分類	多言語化手法
TTS で対応可能	SSML の say-as タグで指定
固有名詞	SSML の phoneme 要素で IPA 記号を記述
一般名詞	オントロジーにより、概念と各言語での表現を記述

この対話制御中心方式は、大局的な対話の流れは MVC フレームワークに従い、局所的な対話の流れは VoiceXML の FIA(Form

[2] http://www.w3.org/TR/voicexml20/
[3] http://www.saltforum.org/

Interpretation Algorithm)などに従うことで、対話の状態遷移、文法、文生成テンプレートなど必要な記述を全て宣言的に記述可能になっている。また、言語依存のプロパティファイルを肥大化させることになってしまうが、文法に関しても非終端記号に相当する部分だけをJSP に記述し、終端記号や句の構成規則をプロパティファイルに記述すれば、ある言語での文法拡張が他の言語でも正しく反映されているかどうかのチェックが行えることになり、拡張作業が確実に行えることになる。

6. おわりに

我々の提案は標準技術とそれを利用したフレームワークを中心的に用いることで、宣言的な知識記述で対話システムが構築できることを目指すものである。インターネットに関する標準技術を中心に対話システムを構成する研究としては ISIS[9]がある。アーキテクチャとしては CORBA を利用した分散アーキテクチャであり、KQML を用いて知的エージェントとのインタラクションを行っている。ISIS はタスクを株式取引に限定し、多言語化よりは、高度なインタラクションの実現を目指している。

提案方式の問題としては、読みの扱いがある。特に固有名詞に関しては認識辞書に本来の言語の音素表記を用いても異なる母語話者が其の通り発音することは期待できない。大半はその母語特有の発音になると考えられ、それらを全て予め準備するのは難しいと考えられる。

参考文献

[1] H. Aust et al.: “The Philips Automatic Train Timetable Information System”. Speech Communication 17, pp. 249–262, Nov. 1995.

[2] 駒谷和範, 上野晋一, 河原達也, 奥乃博.ユーザモデルを導入したバス運行情報案内システムの実験的評価.情報処理学会研究報告, SLP-47-12, 2003.

[3] Qiang Huang, Stephen Cox, “Automatic Call-routing without Transcriptions”, in Proc. EuroSpeech,Geneva, 2003.

[4] J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue, "Multilingual Spoken-Language Understanding in the MIT Voyager System," Speech Communication, Vol. 17, No. 1, pp. 1-18, March 1995.

[5] Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P. and Zue, V.: Galaxy-II: A reference architecture for conversational system development. Proc. of ICSLP 98, 1998.

[6] M. Nakano, T. Minami, S. Seneff, T. J. Hazen, D. Scott Cyphers, J. Glass, J. Polifroni, V. Zue, "Mokusei: A Telephone-based Japanese Conversational System in the Weather Domain," Proc. Eurospeech 2001, Aalborg, Denmark, September 2001.

[7] Yunbiao Xu, Masahiro Araki, Yasuhisa Niimi: A mutilingual-supporting dialog system across multiple domains, Journal of Acoustical Science and Technology, Vol.24, Np.6,pp349-357,2003.

[8] Yan Ming Cheng, Chen Liu, Yuan-Jun Wei, Lynette Melnar, Changxue M: An Approach to Multilingual Acoustic Modeling for Portable Devices, Eurospeech 2003, pp.3121-3124, 2003.

[9] Meng, H. et al., "ISIS: A Trilingual Conversational System with Learning Capabilities and Combined Interaction and Delegation Dialogs," Proceedings of the National Conference on Man-Machine Speech Communication (NCMMSC6), Shenzhen, November 2001.