日本語対話処理研究に必要なもの

伊藤克亘　秋葉友良（電総研）　岡登洋平（筑波大）　河原達也（京都大）
email: {kito,t-akiba}@etl.go.jp, okato@milab.is.tsukuba.ac.jp,
kawahara@kuis.kyoto-u.ac.jp

音声認識・音声合成技術の発展をうけて、わが国においても、1990年代はじめから、数多くの機関によって音声対話システムの研究・開発がすすめられてきた。その結果、これまでに数多くの音声対話データが収録され、多くのデモシステムが試作され、様々な研究が行なわれた。このように多様な研究が行なわれてきたことは、意義深いことである。

しかし、これらの研究は、ほぼ全てが機関もしくはプロジェクトごとの個別の研究に終始しているといっても過言ではなく、個々の研究で得られた知見を他の研究で利用できるどころか、公開されているデータの共通利用すらすすんでいないのが現状である。

本稿では、このような現状把握に基づき、日本語音声対話処理技術を発展させるためには何が必要かを議論するために、4つの視点から問題提起を行なう。

1　自然言語処理の「ツール」を音声研究者はどのように使うべきか（伊藤克亘）

ここでは、自然言語処理の分野で作られてきた日本語処理用の形態素解析システムや辞書・文法を音声認識のために利用するときの問題点について議論する。

1.1　日本語テキストの特徴

まず、日本語の辞書や文法を利用する場合に直面するのが、標準的なものが存在していない点である。これは以下の日本語の特徴に関係している。

- 日本語のテキストは平仮名・片仮名・漢字・その他の記号と、多様な字種からなる
- 日本語には正書法がない
- 日本語には語を認定する標準的な規則がない

これらの要素が組み合わさって形態素の定義に様々な変化を生じさせるため、形態素解析結果も（誤りも含めて）システムごとに様々な変化を生むことになる。

したがって、異なる音声認識システムを評価するときには難しい問題が生じてしまう。

では、SQALE でのドイツ語のように被覆率を目安にすればよいと思うかもしれないが、それだけでは不十分である。ドイツ語では複合語が一語として登録されることはあっても、分かち書きを基準としているため、誤った分割は存在しない。

しかし、日本語の形態素解析の場合、誤った分割が相当量存在する。フリーでない形態素解析システムでは、非常に精度の高いシステムが存在するという噂もある。残念ながら、著者は、新聞1年分程度の形態素解析結果で解析精度が 100% 近いものなどこれまでに見たことはない。誤った分割のうち、一文字からなる短い断片として誤ってしまう場合には、高頻度語に誤りが含まれる可能性が多くなり、被覆率の評価には注意を要する。

1.2　日本語の話し言葉の特徴

日本語では、書き言葉と比較した場合、話し言葉には以下のような特徴がある。

- 文の定義が明確でない（省略、倒置）
- 自分や相手の発話をよく参照する（同じ語の繰り返し、指示語）
- 相手との関係から生じる表現を用いる（待遇表現）
- 相手を意識した表現を用いる（終助詞、間投詞）
- その他（女言葉・男言葉・若者言葉、縮約、漢語・修飾語を余り使わない）

形態素解析システムの辞書や文法は、書き言葉を想定して構築されているため、とくに上記の現象に関係する付属語に関する規則の整備が非常に遅れている。

これでは、対話書き起こしコーパスから対話システム用言語モデルを構築しようと思っても、ある意味で最も対話らしい部分の解析ができずに終わってしまうのではないだろうか。

実は、書き言葉中心の弊害は、品詞の立てかたや辞書の項目の立てかたにも現れている。人間用の辞書にせよ、形態素解析の辞書にせよ、辞書の項目は、漢字表記を中心に構築されている。テキストを処理する場合には、処理のプリミティブは文字なので自然な選択であるといえる。しかし、音声認識から言語をながめた場合、処理のプリミティブとなるのは音韻である。

また、書き言葉にはもともと音声として再生されることを意識しない表現も多くあるため、辞書項目に読みの情報が不足していることも多い。日本語の形態素の読みについては、前後に接続する形態素によって変化する場合がある（連濁など）が、その変化に関する情報は欠落していることが多い。

1.3　音声処理に親和性の高い体系立て

以上のことをまとめると、音声処理のための自然言語処理ツールは、以下のような点に留意して整備する必要がある。

- 文脈独立な読みを区別することを重視した語の標準的な認定規則

- 話し言葉独特の表現に関する規則の整備

これらの整備が実現すれば、形態素解析に利用できるだけではなく、音声認識システムで直接利用できる文法・辞書の整備も可能になる。音声対話システムでは、扱えるドメインを限定するのが現実的であるので、ボトムアップな統計的言語モデルよりも、トップダウン的な文法に基づく設計が可能になることも大きなメリットとなるのではないだろうか。

2 音声「対話」システムのためのツール (秋葉友良)

近年、音声認識技術の高度化、計算機環境の整備、共有資源の増大から、比較的容易に音声対話の実験システムが構築されるようになってきた。しかしそれでも、システム全体の構築にかかるコストは大きく、音声対話研究の進展にとって無視することはできない。ここでは、音声対話システム構築に利用可能な資源のうち、「対話」を実現するために必要なツールについて議論する。

2.1 「対話」に必要なもの

まず、機械との「対話」を実現するために必要となるものについて考察する。音声対話の仕事は、入力に対して適切な出力を返すことである[1]。しかし一般には、その過程が一つのモデルで扱うには複雑であることから、入力と出力を直接に対応付けるモデルを用意することは稀である。また、目的指向対話では、対話の目的 (タスク) を実現する問題解決機を、変換の過程に組み込まなければならない。そのため、一般に次のようなモデルを組み合わせてシステムを構成することになる。

理解モデル 認識結果を問題解決機の入力へ変換
生成モデル 問題解決機の出力を発声文字列へ変換
対話モデル 妥当な入出力のシークエンスのモデル

これらのモデルを構築するための公開されたツールは、決して十分であるとは言えない。むしろ「使える」ツールは少ない。例えば、理解モデルに関するツールは、その中でも比較的多くのツールが公開されている。しかしそれらのうちの多くは、形態素解析、構文解析のツールであり、理解モデルをさらに分割したモデルの一つを実現しているに過ぎない。これらのモデルは、その結果である単語列 (形態素解析) や構文木 (構文解析) が、問題解決機への入力生成あるいは対話モデルで扱う発話単位の生成に「役に立つ」と判断されれば利用するに値する。さらに抽象度の高い[2]表現を生成するモデルとして、「意味解析」が考えられるが、その結果となる意味表現に一般的なコンセンサスがないため、また高度な意味表現抽出にはモデル独自の知識 (文法、辞書) が必要となるため、広く利用できるツールはほとんど存在していないというのが現状である。また、実際に高度な意味表現を生成する必要があるかどうかは、問題解決機で扱うタスクの複雑さに依存するため、一概に論ずることはできない。いずれにせよ、ツールを使って構文木や意味表現などが得られた後、そこから問題解決機への入力となる (問題解決機固有の) 表現を抽出してやる必要はある。すなわち、独自の問題解決機を使用する限り、理解モデル中に問題解決機に依存した (したがって、既存のツールで解決できない) 部分が何らかの形で組み込まれることになる。

[1]入力と出力が一対一に対応して交互に現れるとは限らない。

[2]本節では、入出力となる生の表現 (文字列) と比べて、より多くの操作 (情報の付加・簡素化・変形等) が加えられた表現を、より抽象度が高いと呼ぶこととする。

2.2 「対話」ツールの必要性

音声対話システムの構築に使えるツールが少ないという現状では、まずシステム全体の構築にかかるコストが大きくなるという直接のデメリットの他にも、次のような間接的なデメリットがある。各システムで独自のインプリメントを行わざるを得なくなるため、システム内部の透過性が悪くなり、他システムとの比較が容易ではない。また、一部のモデルだけを取り出した実装や評価が行い難い。

では、現在なぜ使えるツールが少ないか、なぜツールが利用されないのであろうか。第一の理由は、モデル間のインターフェースの問題である。これは自然言語処理研究一般の問題でもある。一般に、抽象度が高い表現を出力するモデルほど、結果の形式の自由度が高く、モデル間で統一が取りにくい。表現形式の統一化には、モデルの分割区分 (構文解析と意味解析に分ける、など) の検討、共通に利用できる表現の検討が必要であり、なかなか難しい課題ではある。しかし現時点では、米国の ATIS などのように共通のタスクを設定した上で、少なくとも問題解決機の入出力の形式を (例えば、SQL などに) 統一し、その理解モデル、生成モデルに必要なツールを整備していく必要があると考える。

一方、対話モデルについては、前述の 3 つのモデルのうちでも、特に研究を進める必要があると筆者は考える。本格的な対話モデルの研究は長期を要するものであるが、なるべく早い段階から設計、実装、評価のサイクルを繰り返すことが重要である。対話モデルの理論研究では、発話行為という抽象度の高い表現のシークエンスで捉えるものが主流である。しかし、音声対話という、ヒューマンファクターまでを踏まえる必要のある対話を、そのような抽象的なレベルで扱ってよいのか、扱えるのかということは、全く明らかになっていない。音声対話研究として今現在必要なのは、発話された文字列とそうかけ離れていない、抽象度の低い表現のシークエンスを、オートマトンなどの簡単な原理でモデル化し、設計・実装・評価サ

イクルを容易に実行できる枠組みであると考える。

3 音声対話コーパスの現状 (岡登洋平)

これまでに構築された音声対話コーパスを概観し、これらのコーパスを構築・利用する上での問題点について議論する。

3.1 音声対話コーパスの利用

音声対話処理研究を進めていく上で音声対話コーパスは非常に重要である。日本語の音声対話についても、いくつかのコーパスが公開されている。それぞれの対話コーパスは、これまで主に単独のコーパスとして研究に利用されてきた。これらのコーパスは作成された目的や時期が異なり、コーパスの大きさ、扱う対話のドメイン、利用できる情報も異なっているものの、全体で 385 対話、約 19 時間の対話データにのぼる。今後、実用的な音声対話処理を行なう上で、対象となるドメインを考慮してモデルをチューニングする必要は当然あるにしても、音韻モデルや言語モデルの構築にこれらのコーパスを利用することができれば、研究はより加速すると考えられる。またそのように利用しにくいのであれば、コーパスを構築する際の問題として、検討する必要がある。そこで、コーパスを同時に利用するにあたり、問題となる点について考えてみたい。

3.2 音声対話コーパス利用時の問題点

これまでに公開されている音声対話コーパスを表 1 に示す。ADD は ATR 対話データベース、ASJ は音響学会連続音声データベース Vol.7、PASD は文部省重点研究音声対話データベース、RWC は RWC 音声対話データベース、をそれぞれ表す。

表 1 日本語音声対話コーパスの比較

	対話数	音声データ		書き起こし	
		時間 (分)	時間情報	文数[3]	文字数[4]
ADD	262			24K	585K
ASJ	37	236 × 1	×	4K	95K
PASD	69	404 × 2	△	6K	114K
RWC	48	525 × 2	○	11K	199K

これらのコーパスを用いる上でまず問題なのは、評価に用いる具体的なタスクを想定した音声対話コーパスが存在しないことである。書き起こしテキストは全てのコーパスに付与されているが、より現実的な正解を準備する必要がある。今後の整備が期待される。

またそれぞれのコーパスで利用できる情報が異なる。書き起こしと読み情報は全てのコーパスで利用できる。音声情報は、ASJ ではモノラルで収録されており、PASD, RWC ではステレオで話者ごとに分離されて収録されている。PASD の一部や RWC は主にポーズで区切られた音声を単位として、時間情報と発話内容が付与されている。

今後、当面は音声データ (話者ごとに音声を分離)、発話内容 (書き起こし、読み)、時間情報が主に利用できると考えられるので、これまで収集されたデータについても、不足している情報を付与して公開すれば、これらのコーパスの有用性が増すはずである。

同じ音声情報や書き起こしでも、音響的性質、音声的、言語的な性質が収録サイトやコーパスごとに異なりがある。対話の収録は収録環境の違いが大きく影響する。また書き起こしも作成者により傾向が異なり、全体として、サイト間やコーパス間で偏りを生じる可能性がある。例えば「いう」と「ゆう」という表記は、ADD, ASJ, PASD では「いう」がほとんどの場合使われているが、RWC では「ゆう」が多く使われている。このような違いは統計的な処理を行なう際、注意する必要がある。コーパスの均質性は重要だが、データの収集上のばらつき自体、重要な問題であり、研究に重要な支障を生じるのでなければ、ある程度のばらつきは始めから考慮し、性質を調べたりツールなどにより吸収する方が望ましい。このような問題点やコーパスの誤りの修正は、コーパスを利用する過程で蓄積されていくものであり、公開後もコーパスの保守を続けていく必要がある。

今後、大規模なコーパスを構築していくには、サイト間の協力は避けられない。これまで複数のサイトで収集したコーパス ASJ, PASD はサイトごとに異なる対話のドメインを設定している。これらのコーパスは、利用に際して自サイトで収集したデータのみが使われる傾向がある。これは各サイトが利用したいドメインのデータを収集したこともあるが、全体的として、それぞれのドメインについて公開された対話数が少ないこと、様々な面でデータの素性が異なっているために利用しにくいことも原因だと思われる。そのため、利用しやすいコーパスを構築するには、複数のサイトで同一のドメインの対話データを収集する必要がある。複数サイトで同一のドメインの対話を収集することで、利用できるデータが増えると同時に、サイト間のばらつきにも対応しやすくなると思われる。

今後コーパスを構築していく上で標準的な解析体系、ツールの整備が欠かせない。これまでに収集されたコーパスで共通して利用できる情報は音声データ、書き起こしテキスト、一部の発話情報などに限られる。これは解析方法やラベル付けの体系が標準化されていないことが問題である。例えば ADD は書き起こしの他に品詞や構文の情報が含んでいるが、ADD に付与されているものと同じ品詞体系で他のコーパスに品詞を付与するのは困難で、利用は難しい。発話単位の設定に関しても同様の問題がある。標準的な

[3]文は句点ごとに区切り、間投詞は単独で文として計算した。
[4]文字は音声の書き起こし部分のみを数えたものである。

ツールが整備されていけば、現在よりも容易かつ高度な利用が可能になると思われる。

4 音声言語理解研究のためのリアルな評価用データベースの必要性 (河原達也)

音声対話システムのための音声言語理解の研究に必要なデータベースに焦点をあてて議論する。

4.1 大規模なデータの必要性

音声対話システム及び音声認識を伴う音声応答システムは、音声認識技術の最も有望なマーケットの一つであると考えられる。我が国においても、約5年前から数多くの機関で研究・開発が行われ、デモシステムも作成された。しかし著者の知る限り、フィールドテスト（多くの一般の人に試用してもらい評価を行うテスト）が実施されるに至ったものはそのうちのごく少数であり、実際に運用されたのは皆無である。仕様の段階において、実際の応用としての有用性を無視していたり、あるいは（今後10年の技術革新を見越しても）実際の使用に耐えないことが明白であるようなデモシステムも少なくない。これらは、多様な研究が行えた点で有意義であったが、技術的には何ができて何が（重要な未解決の）問題であるかを十分に明確にできないという結果となった。[5]

欧米においては、純粋な研究プロジェクトとしては、ARPA主導のATIS（フライト情報検索）やESPRIT主導のKIOSK（列車情報検索）のように、共通のタスク設定のもとで大規模なデータ収集が行われてきた。これらのタスクドメインに共通するのは、アプリケーション（ニーズ）及び技術（シーズ）の両面において（スポンサーと研究者が納得できる）現実味が高いことである。もちろんこれらには、シナリオが用意されたものでデータも人為的であるとか、実際にシステムを構築する際にこれだけの大規模な（学習用）データを用意できないといった問題点もあるが、研究のためのインフラストラクチャーとしては十分であった。

これらの成果(?)に基づいて、米国のいくつかの企業は音声言語システムのプラットフォームを販売しているが、これも実際に大規模に運用されているという例を著者は知らない。これは実際のアプリケーション（とヒューマンファクタまで）を考慮してシステムを構築することが成功に必要であることを示唆している。

著者が滞在していたAT&Tでは頻繁にフィールドテストを実施していた。タスクドメインの設定にあたっては、マーケット及び技術の両面から綿密な検討を行っている。個々のシステムに関して、数サイクルにわたってフィールドテストを実施し、各サイクル毎に評価及びシステムへのフィードバック（チューニング等）を行う。各サイクルでは数百～千発話程度（話者も百名程度）のサンプルを収集する。これは、ある程度のユーザや発話の現象を包含した上で評価を行うために必要なデータ量を示唆している。[6]

4.2 リアルな人間－機械対話データの必要性

音声対話システムの研究が進展するにしたがって、実際にユーザに使用してもらうには、種々の音声や言語上のバリエーションに対処する必要があり、それは容易ではないことが明らかになってきた。一方、一般のユーザが機械に話しかける際には人間（のオペレータ）に話しかけるのとは異なることもわかってきた。したがって、リアルな人間－機械の対話のデータを収集することが重要であると考える。[7]

WOZ方式は一つの方法ではあるが、WOZの役割と能力を明確に定義しておくことが不可欠である。少なくとも発話を整形して入力すれば機械的に処理・応答生成を行えることが必要である。[8]そういう意味で、著者はできるだけ簡単なタスクがよいと考える。できればWOZではなく、音声認識器を用いて収集するのが望ましいと考える。実際に音声対話システム全体の評価にはタスク設定やヒューマンインタフェースが少なからず影響するが、音声言語理解の研究のためには、これらの要因を考えなくてすむような簡単なタスクが望ましい。

タスクとしては簡単でも、実際の応用として有用であることは不可欠である。すなわち実際にサービスが提供されれば（コストは無視しても）使ってみたいと思うようなタスクでないと、リアルなデータは収集できない。できればシナリオを用意しなくても、発話できるようなものが望ましい。

4.3 まとめ

以上の議論をふまえて、リアルな評価用音声対話データベースの要件を以下にまとめる。

- 機械で処理できるタスク設定
  問題を音声言語の側面に絞り、機械相手の現実的な発話を収集
- 実際にサービスとして成立しうるタスク設定
  シナリオなしに自発的な発話を収集
- 多数（100名程度）の話者による、多数（1000発話以上）のサンプル
  種々の発話の現象を包含

具体的なタスクについては、発表（討論）の際に提示したいと考えている。

---

[5] 音声認識システム自体もはや一般の人にとって夢でなく、単にトイシステムを作ればよいという段階ではない。また音声や対話のもっとスパンの長い基礎的な研究とは問題を区別すべきである。

[6] 音響モデル構築のための話者数や、認識実験のためのサンプル数にもほぼ同様の議論（尺度）が成立すると考えられる。

[7] もちろん対話の分析やモデル化を行う上では、人間－人間の対話データは必要であることは言うまでもない。

[8] 今後何十年かの人工知能研究を待つのであれば別であるが。