ポータビリティを指向した音声対話データベース検索システム

駒谷 和範　　鹿島 博晶　　安達 史博　　河原 達也

京都大学 情報学研究科 知能情報学専攻

komatani@kuis.kyoto-u.ac.jp

概要

音声対話システムを設計するにあたってシステムのポータビリティは重要な要素である。我々はフレーズ文法と汎用的な統計的言語モデルを融合した言語制約を用いることで、タスク遂行に必要なキーフレーズをスポッティングする手法を実現した。また語彙及び意味解釈機構といった、ドメインデータベースから得る知識の抽出法を改善した。ホテル検索タスクにおける評価実験の結果、提案したキーフレーズスポッティングにより意味理解誤り率は15.5%削減された。

1　はじめに

現在音声対話システムの研究がさかんに行われつつあるが、不特定多数のユーザを対象として広く実用的に使われているシステムは見当たらない。音声認識誤りへの対処が不十分であることとともに、一度構築したシステムを他に移植することが困難であること（ポータビリティの低さ）が原因として挙げられる。

現在大語彙音声認識ではN-gramに基づく統計的手法が成果をあげているが、これを音声対話システムで使用する場合には、タスクに合致したデータを大量に集める必要が生じる。さらにタスク遂行に必須のキーワード部分（固有名詞など）を正確に認識しなければならないが、統計モデルの場合では、頻度の低い固有名詞やコーパスにあまり現れないキーフレーズを確実に認識することは困難である。また大量のコーパスを収集したとしても、特定のシステムのために収集されたコーパスはポータビリティが低くそのままでは再利用が難しい。

そこで本研究室では、検索対象となるデータベースからの情報をもとに、文法と語彙の設定を半自動的に行う汎用的な情報検索音声対話プラットフォームを作成した[1]。これは検索に必要な語彙をデータベースから取り出したうえで、典型的な発話パターンを文法として半自動的に生成するものである。しかし文法をベースに音声認識を行うため、生成した語彙や文法から外れた非定型な発話にうまく対処することができなかった。

そこで我々はタスクの汎用性を考慮しながら多様な表現を受理可能とするために、統計モデルと記述文法を複合的に用いた言語制約によって、タスク遂行に必要なキーフレーズをスポッティングする手法を実現した[3]。また実際のユーザ発話を分析した結果を用いて、音声対話システムに必要な語彙及び意味解釈機構といった知識を、ドメインデータベースから抽出する方法を改善した[2]。このような音声対話システムにおいて、音声認識結果の信頼度を利用して効率的な確認や誘導を行う[4]ことで、ポータビリティを備えつつ、より多くのユーザの発話に対処可能なシステムとなる。以降、2章では音声認識部に関して、3章で解釈部に関して、4章で対話管理部について述べる。

2　複合的言語制約に基づくフレーズスポッティング

近年の音声対話システムでは、そのタスクに特化した対話コーパスが大量に得られる場合、N-gramモデルが用いられている[5]。この言語制約により、非定型な発話に対しても柔軟に対処できるが、個々のタスク毎に大量の対話コーパスを収集するのは困難である。

このことから記述文法[1]を言語制約に用いたシステムも多い[6][7]。[8]ではシステムの移行に際し、元のタスクの2-gramを用いることができず、新たなシステムでは記述文法を用いている。記述文法の利点として、タスクに特化した知識を容易に導入でき、また語彙などの変更も簡単に行えるという点がある。しかし、受理可能な発話を完全に限定してしまうため、ユーザに自由な発話を許さないという問題がある。さ

[1]ここでは、文脈自由文法や有限状態文法などのルールベースの文法を指す

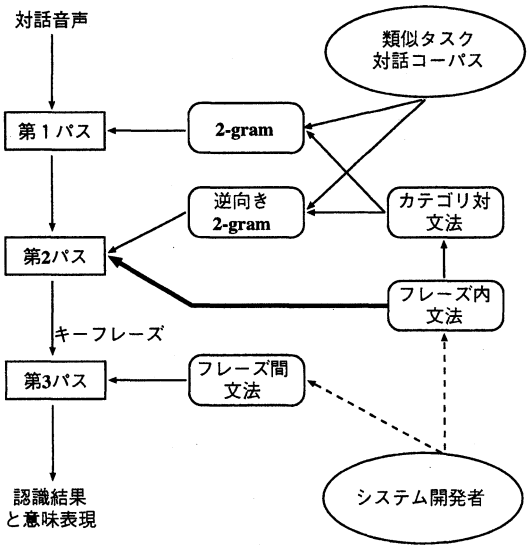

図 1: 複合的言語制約を用いたフレーズスポッティング

らに、全く異なったタスクへの移行の際に、文単位の文法を全て作成し直すには、大きなコストがかかる。

そこで我々は、タスクに特化したキーフレーズ部分に記述文法を、タスクに依存しないフィラー部分[2]に対しては類似タスクのコーパスから得られる 2-gram を与える、複合的言語制約に基づくフレーズスポッティングを実現した [3]。キーフレーズとフィラーが任意に接続するような緩い言語制約に基づく従来のスポッティング手法と比較して、フィラー部分に単語 N-gram を適用することにより文全体に対する言語制約が強くなり、認識精度は向上する。また、この単語 N-gram は類似タスクの大規模コーパスを利用でき、タスクの完全な合致を要求しないので、タスクのポータビリティもよい。さらに文全体を N-gram でデコードするのではなく、意味理解に必要なキーフレーズに着目してスポッティングすることにより、多様なキーフレーズ候補を効率よく求めることができる。

従来のスポッティングを用いた手法では、その単位として主にキーワードを用いることが多い。しかし、単語のテンプレートでは局所的な類似性やノイズの影響を受けやすく、より長いフレーズを単位とした方が抽出精度が高いことが示されているため、ここではフレーズをスポッティングの単位とする。フレーズは、「所在が」や「一万円以下の」のようにキーワードとその付属語からなり、対話音声のような非定型な発話においてもその構文が保持される。さらに、フレーズの単位は意味表現にも直結するため頑健な音声理解が期待できる。

[2]キーフレーズ部分以外をフィラー部分とする

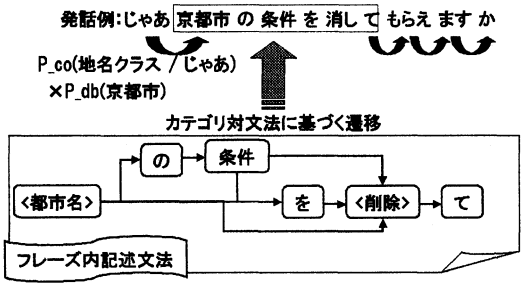

図 2: 構築した 2-gram の適用例

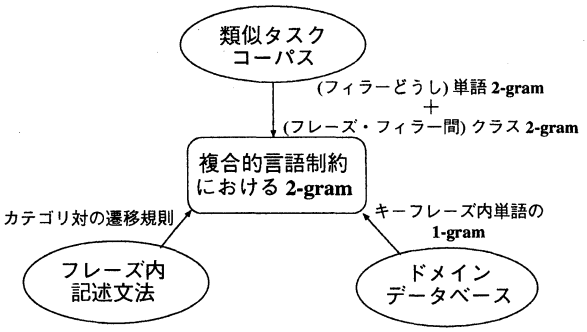

図 3: 2-gram 構築における複合的言語制約の適用

本手法では、特にキーフレーズに着目した段階的探索を行う。これは、以下のようにキーフレーズ部分に対し、段階的に強い言語的制約を適用していくものである。

第 1 パス 単語 2-gram モデル

第 2 パス スポッティング時に局所的にフレーズ内文法

第 3 パス フレーズ接続時に意味制約としてフレーズ間文法

第 1, 第 2 パスにおいて、単語 2-gram と記述文法の複合的言語制約に基づきキーフレーズをスポッティングする（図 2）。フィラー・キーフレーズ間やフィラーどうしの言語制約として単語 2-gram を用い、キーフレーズ内の言語制約には記述文法を用いる。また、

表 1: 対話音声理解を目的としたFA+SErrによる従来方式との比較

	語彙数	文法内発話	準文法内発話	文法外発話	total
正解数		561	116	120	797
文単位 記述文法	942	14.8%	51.4%	175.2%	45.8%
フレーズスポッティング (連接モデル)	1290	16.8%	34.2%	154.2%	37.9%
フレーズスポッティング (複合的制約)	6124	**10.8%**	**24.7%**	**140.9%**	**30.3%**

FA:誤受理率 (= 1 − 適合率), SErr:スロットエラー (= 1 − 再現率)

表 2: 検索条件指定発話の内訳

	発話数
システムの想定内の発話	348
語彙外の発話	168
文法外の発話	20
タスク外の発話	374
合計	910

表 3: タスク外の発話の内訳

タスク外発話となった原因	発話数 (内訳)
DB にフィールドがない	130 (34.8%)
DB にフィールドはあるが値がない	133 (35.6%)
DB に値はあるが修飾語がついている	6 (1.6%)
検索要求が正しくない	1 (0.3%)
数値の程度を表す形容表現を含む	12 (3.2%)
値に漠然とした範囲を指定している	61 (16.3%)
検索要求以外	27 (7.2%)
呼び掛け、無関係な発話	4 (1.0%)
合計	374 (100%)

ドメインデータベースから得られる 1-gram を用いてキーフレーズ内の単語に出現頻度を付与する (図 3)。

第 3 パスでは、スポッティングされたフレーズを組み合わせて、文として認識・理解する。この場合、フレーズ候補はそのスコアと意味制約にしたがって接続される。意味制約はシステム開発者が記述するフレーズ間文法を用いる。ただし、このままでは単語数の少ない仮説が優先的に受理されることになるため、フレーズ間のスキップする区間に比例した一定値を評価関数から減じながらフレーズを接続する。

話者 24 名 (男性 19 名、女性 5 名) に対する評価実験の結果、文全体を記述文法で表現した言語制約と比較して、文法内、文法外いずれの発話においても誤り率の減少が見られ (表 1)、全体では 15.5%の誤りが削減された [3]。

3 汎用性を考慮した解釈機構の拡張

ポータビリティを考慮して音声対話システムを構築する際には、認識に必要な語彙や解釈に必要な規則をドメインデータベースから抽出することが効率的である [1]。しかし、より多くの表現に対応するためにはドメインデータベースから抽出する語彙や文法を拡張することが必要となるが、むやみに語彙や文法を追加することは認識性能の低下につながる。

そこで、データベースからそのまま半自動的に語彙と文法を生成したシステム [1] を初心者 24 名が使用した結果を分析し、実際に起こったシステムの想定外の発話 (語彙外・文法外・タスク外) への対処となるように、語彙や解釈規則の抽出法を改善した [2]。その結果を以下に示す。

省略語の補完

収集した対話データには、所在の地名を表すのに「府」「県」「市」を省略した発話が多くあった (表 2 の「語彙外の発話」中の 77 発話)。そこで「府」「県」「市」を省略した地名も語彙として抽出したうえで、データベース内にある地名と照らし合わせて省略を補完できるように意味解釈規則を用意した。「京都」のように「京都府」「京都市」の二通りに補完可能なものは、「京都は京都市ですか？京都府ですか？」のようにユーザに聞き返すことで曖昧性を解消する。このような補完は、日付の表現において「1997 年」を「97 年」と表現するような場合にも適用できる。

内容語に付加される接尾辞

「京都市」に対する「京都市内」など、内容語に付加される接辞を語彙及び文法に追加する。検索時にはこれを取り除いて検索する。

数値の程度を表す形容表現

数値の程度を表す形容表現は、料金フィールドなら「安い」「高い」、時間フィールドなら「早い」「遅い」のようにフィールドから予想できるため、それらを語

彙に登録する。値はフィールドごとにシステム開発者が与えておく。この値はタスクやフィールドに依存した知識であるため、システム開発者が与えるのが妥当である。

漠然とした値の範囲を示す表現

数値の漠然さを表す表現も語彙に登録する。これは数値を表すフィールド一般にあり得る表現である。例えば、料金フィールドなら「ぐらい」「前後」などが予想できる。これもシステム開発者が値の幅を与えておくことで解釈および検索が可能となる。

以上のように改善した抽出法に基づいて語彙と解釈規則を再構成し、検索・応答が可能となることを確認した。また、これらを実現することで、想定外であった発話のうち約46%に対して正しい応答を行うことができるようになった[2]。

4 信頼度を用いた対話管理

「よく聞き取れなかった言葉について確認を行なう」ということは、人間同士の対話でもよく行なわれることである。そこで、音声認識器のN-best出力(N=10)とそのスコアからキーワードに対する事後確率を計算し、その単語の信頼度とした。これは直観的には、どの文候補にも一貫して出現する単語は信頼できるとみなすものである。具体的には第 i 候補の対数尤度を $logscore_i$ として、単語 w の信頼度を以下のように定義する。

$$p_i = \frac{e^{\alpha \cdot logscore_i}}{\sum_{j=1}^{N} e^{\alpha \cdot logscore_j}}$$

$$CM_w = \sum_{i=1}^{N} p_i \cdot \delta_{w,i}$$

ここで、単語 w が第 i 候補に含まれるとき $\delta_{w,i} = 1$, 含まれないとき $\delta_{w,i} = 0$ である。α はスムージング係数で、予備実験の結果 $\alpha = 0.05$ とした。

単語（＝キーワード）レベルで認識結果に信頼度 CM_w が付与されると、以下のような確認戦略が可能になる。一発話中に複数のキーワードがある場合は、それぞれについて、受理／確認／棄却の判断を行う。

- $CM_w \geq \theta_1$ → 確認なしで受理する
- $\theta_1 > CM_w \geq \theta_2$ → 直接的に確認を行う
  「○○でよろしいですか？」
- $\theta_2 > CM_w$ → 棄却する

また、概念レベルでも発話内容の意味カテゴリについて信頼度を計算することにより、単語レベルで十分に信頼度が高くない候補が得られなかった場合においても、適切な発話の誘導を行うことができる[4]。

5 まとめ

本研究室で開発しているデータベース検索音声対話システムの構成について述べた。音声認識部では汎用性に優れた複合的言語制約に基づくフレーズスポッティング手法を実現した。また、収集した対話データに基づいて、音声対話システムに必要な知識をドメインデータベースから抽出する方法の改善を行った。

今後、今回実装したシステムを用いて再度被験者によるシステム全体としての評価を行う予定である。また不特定多数のユーザが使用可能なシステムを目指してさらに改善をすすめていく。

参考文献

[1] 田中克明, 河原達也, 堂下修司: 汎用的な情報検索音声対話プラットフォーム, 電子情報通信学会技術研究報告, SP98-109, NLC98-45 (98-SLP-24-14) (1998).

[2] 安達史博, 駒谷和範, 河原達也: 音声対話情報検索システムにおける想定外の発話の分析とその対処, 人工知能学会研究会資料, SIG-SLUD-A001-2 (2000).

[3] 鹿島博晶, 河原達也: 複合的言語制約に基づくキーフレーズスポッティングによる対話音声理解, 電子情報通信学会技術研究報告, SP2000-114, NLC2000-66 (SLP-34-40) (2000).

[4] 駒谷和範, 河原達也: 音声認識結果の信頼度を用いた頑健な混合主導対話の実現法, 情報処理学会研究報告, SLP-30-9 (2000).

[5] San-Segundo, R., Pellom, B., Ward, W. and Pardo, J.: Confidence Measures for Dialogue Management in the CU Communicator System, *Proc. ICASSP*, Vol. 2 (2000).

[6] 中野幹生, 堂坂浩二, 宮崎昇, 平沢純一, 田本真詞, 川森雅仁, 杉山聡, 川端豪: TV 番組の録画予約を受け付ける実時間音声対話システム, 情報処理学会研究報告, 98-SLP-22-8 (1998).

[7] 桐山伸也, 広瀬啓吉: 文献検索音声対話システムの機能拡張とその評価, 情報処理学会研究報告, SLP-30-10 (2000).

[8] 小暮悟, 伊藤敏彦, 中川聖一: 音声対話システムの移植性に関する考察-観光案内システムとデー タベース検索システム-, 情報処理学会研究報告, SLP-25-3 (1999).