キーワードと汎用言語モデルに基づく統計的音声対話システムにおける音声認識部・応答選択部の密結合

学籍番号　２３４１３５５９　氏名　平野 隆司
指導教員名　李 晃伸

1　はじめに

本研究では，ユーザ生成型のようなキーワードに基づく音声対話システムを想定し，その高精度化について述べる．ユーザ生成型音声対話コンテンツ [1] ではユーザが対話コンテンツを登録してもらうことで，ユーザによる対話データベース構築が可能である．しかし，対話コンテンツとして質問キーワードと対応する応答文の組が登録されるため，音声認識部と応答選択部はキーワードのみで結合されていた．しかし，音声認識で得られる情報にはキーワード以外にも応答選択に有用な情報が含まれていると考えられる．

そこで，キーワードの情報に加え，汎用言語モデルの情報を用いてシステムの構築をすることにより，音声認識部と応答選択部をキーワードとそれ以外の発話部分によって密結合するシステムを提案する．

2　一問一答形式の統計的音声対話システム

一般的な一問一答形式の統計的音声対話は，ユーザが質問発話した音声の特徴に対して出力確率が最大になるような応答を選択する問題として捉えることができる．質問発話の音声信号系列を入力 O，それに対する応答文を出力 A とするとき，以下のように定式化できる．

$$\hat{A} = \underset{A}{\operatorname{argmax}} P(A|O) \tag{1}$$

質問発話の音声から直接応答文を選択することは困難なので，中間表現に単語列 W を定義することで，以下の式のように置き換えられる．

$$\hat{A} = \underset{A}{\operatorname{argmax}} \sum_{W} P(A|W)P(W|O) \tag{2}$$

式 (2) の $P(A|W)$ は応答選択部，$P(W|O)$ は音声認識部から与えられ，それぞれの統計モデルが適切に学習されることが重要である．そして，真の確率分布を推定するためには，大量のデータを学習する必要がある．このとき，音声認識部 $P(W|O)$ の学習には質問発話に関する音声波形とテキストコーパスが，応答選択部 $P(A|W)$ では質問文と応答文を対応付けた対話データが必要となる．しかし，適切な応答を選択するためには，選択の基準となる有用な情報がより多く含まれていることが重要となる．さらに，その有用な情報はタスク特有の知識である場合が多く，タスクごとにデータの収集が必要となる．このとき，タスク知識の量とデータ収集のコストはトレードオフの関係にある．

タスク知識として発話文集合ではなく，キーワードのみが与えられる場合，式 (1) の中間表現にキーワード K を用いたキーワードに基づく音声対話システムがあり，式 (3) で表される．

$$\hat{A} = \underset{A}{\operatorname{argmax}} \sum_{K} P(A|K)P(K|O) \tag{3}$$

この枠組みでは，タスク知識となりうる単語をキーワードと設定することで，応答精度を保ちながら，データ収集のコストを抑えることが可能である．このシステムにおいて，音声認識部ではキーワードを認識する必要があり，キーワードのワードスポッティング手法が考えられる．式 (3) の $P(K|O)$ では，質問発話の音声信号系列から直接キーワードを求めることでキーワードの抽出精度を高めている．そのため，応答選択に必要な知識を正確に抽出することで，システム全体の応答性能の向上が期待できる．

そして，新たなキーワードに基づく音声対話システムとして，ユーザ生成型音声対話コンテンツ [1] が提案されている．タスク知識をキーワードと応答文と定義し，タスク設定をユーザが行うことで，幅広いタスク知識を含んだ対話データベースを構築可能である．しかし，キーワードに基づく音声対話システムでは，タスク知識をキーワードのみに限定しているため，様々な発話表現や言い回しなど，キーワード以外の発話部分を考慮していない．そのため，発話から応答を返すという観点では，発話に対しロバストな応答選択を行うための知識が不足しているといえ，キーワード以外の発話部分にも応答選択に有用な情報がある可能性がある．

3　音声認識部と応答選択部の密都合

キーワード以外の発話部分を利用して，音声認識部と応答選択部を密に結合するシステムの枠組みを提案する．このとき，タスク知識以外に使用できる情報として汎用言語モデルを考え，応答選択に使用する情報を補強する．キーワード以外の発話部分（ガーベージ）を G と定義すると，キーワードとガーベージに基づく統計的音声対話システムは以下のように書ける．

$$\hat{A} = \underset{A}{\operatorname{argmax}} \sum_{\{K,G\}} P(A|K,G)P(K,G|O) \tag{4}$$

ここで，$\{K,G\}$ は質問発話の書き起こし文を表しており，文 W と同じような内容であるが，$\{K,G\}$ はガーベージ G が与えられない．

応答選択部 $P(A|K,G)$ の学習には，応答文 A と対応する文 $\{K,G\}$ が必要となる．しかし，ユーザ生成型を考えた場合，学習には A と K の組のみしか与えられないため，K から $W' = \{K,G\}$ を生成する必要がある．そこで，汎用言語モデルに基づいてキーワードからそれ以外の発話部分を補間し，文として自動生成する．キーワードの情報を λ_k，汎用言語モデルの言語情報を λ_g とるすと，W' は以下の式のように生成される．

$$\hat{W'} = \underset{W}{\operatorname{argmax}} P(W|K,\lambda_k,\lambda_g) \tag{5}$$

ここで，得られた W' を用いて応答選択部を学習した場合，応答選択部は $P(A|W',\lambda_k,\lambda_g)$ と書ける．

次に，音声認識部 $P(K,G|O)$ は，$W' = \{K,G\}$ と表すと $P(W'|O)$ と表現でき，一般的な音声認識デコーダで処理可能である．しかし，応答選択も含めたシステム全体の性能を考えた場合，タスクに適したキーワードにはより重要な情報が含まれると考えられる．そこで，キーワードの情報を十分に生かしつつ，文 $W' = \{K,G\}$ を出力するような音声認識部が必要となる．そのため，ディクテーション中に動的にワードスポッティングを行う手法を検討する．キーワード情報を用いてキーワードを抽出しつつ，それ以外の発話部分は汎用言語モデルにより補間される．このような音声認識部は，$P(W'|O,\lambda_k,\lambda_w)$ と記述できる．

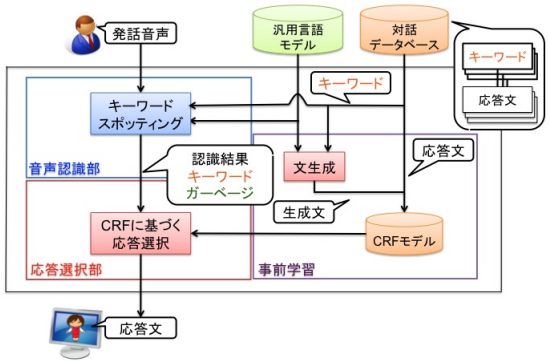

図 1: キーワードと汎用言語モデルに基づく密結合システム

上記の考察を踏まえたうえで，式 (4) のシステムを書き直すと，以下のように表せる．

$$\hat{A} = \underset{A}{\operatorname{argmax}} \sum_{W'} P(A|W', \lambda_k, \lambda_g) P(W'|O, \lambda_k, \lambda_g)$$
$$\text{ただし，} W' = \{K, G\} \tag{6}$$

この式 (6) の提案システムでは，ユーザ生成型の制約から中間表現としてキーワードを用いる必要があるものの，キーワード情報 λ_k と一般的な言語情報 λ_w を十分に各部で利用することで，応答選択部と音声認識部をより密に結合したものと見なすことができる．これにより，少量のタスク知識のみが与えられる条件下でも応答性能の向上が期待できる．

式 (6) に基づいて構築したシステムの全体構成図を図 1 に示す．音声認識部では，ユーザの発話音声に対して $P(W'|O, \lambda_k, \lambda_g)$ が最大となる $W' = \{K, G\}$ を出力する．ここでは，キーワードと N-gram 言語モデルを用いてディクテーション中に動的にキーワードスポッティングを行う [2]．登録されるキーワードは複数であることが想定されるので，セットとしての抽出率を高めるために，複数キーワードの共起制約を用いる．認識した単語列 W' は形態素解析後，応答選択部に渡される．

次に応答選択部では，モデル $P(A|W', \lambda_k, \lambda_g)$ に基づいて認識された単語列 W' 対して確率が最大となる応答文 A' を出力する [3]．応答選択モデル $P(A|W', \lambda_k, \lambda_g)$ は，条件付き確率場 (Conditional Random Fields; CRF) に基づいて構築する．モデル学習では，N-gram に基づいてキーワードから自動文生成を行い，生成文と応答文の学習データベースを作成する [4]．その後，生成文のみ形態素解析し，応答選択モデル $P(A|W', \lambda_k, \lambda_g)$ を学習する．

4　評価実験

一問一答形式の音声対話システム「たけまるくん」のデータベースを用いて評価実験を行った．このデータベースは，質問文と応答文のデータで構成されている．それを元に，質問文からキーワードを抽出することで，キーワードと応答文のデータベースを構築した．汎用的な言語モデルには，Web から収集したテキストコーパスから学習した単語 3-gram モデル (WebLM) を使用する．また，タスク知識が十分にある場合と比較するため，たけまるくん言語モデル（たけまる LM）を用意した．以下に比較したシステムとその構成を示す．

- 発話文システム $P(A \mid W)P(W \mid O)$
  大量の質問文と応答文のデータがある場合に構築可能な理想的なシステム．音声認識部は たけまるくん LM を用いたディクテーション．応答選択部は発話文と応答文を学習した．

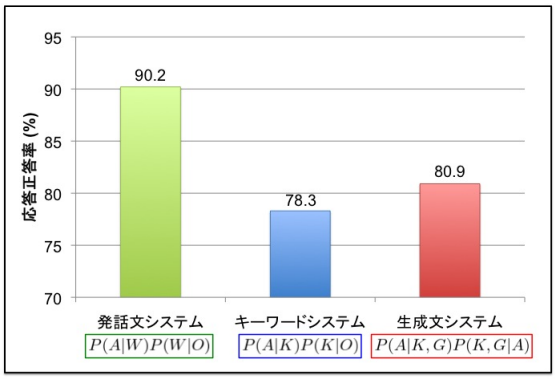

図 2: 各システムの応答正答率

- キーワードシステム $P(A \mid K)P(K \mid O)$
  キーワードと応答文のデータベースを元に構築した従来システム．音声認識部は WebLM とキーワードを追加した Web 辞書を用いたディクテーションと，テキストマッチングによるキーワード抽出．応答選択部はキーワードと応答文を学習した．
- 生成文システム $P(A \mid K, G)P(K, G \mid O)$
  キーワードから生成した文と応答文のデータベースを用いた提案システム．音声認識部は WebLM とキーワードを追加した Web 辞書を用いた共起制約に基づくスポッティング．応答選択部は生成文と応答文を学習した．

図 2 に各システムの応答正答率を示す．発話文システムはタスクに沿った発話文を学習したため，3 手法の中で 90.2% と最も高い応答正答率であった．キーワードシステムではコーパスを利用していないので，11.9% 低下した．その誤りの多くはキーワードの抽出誤りに依るものであった．そして，生成文システムでは生成した文を学習することで，2.6% の改善が得られた．そのサンプル数は 38 個であり，多くがキーワードを正確に抽出できたことで改善がみられた．

さらに，音声認識部をスポッティングに固定し，応答選択部のみの比較を行った．このとき，キーワードのみを学習したシステムは 79.3% であり，提案システムの方が 1.6% 高く，ガーベージを利用することの有効性が確かめられた．

5　むすび

本研究では，キーワードに基づく音声対話システムを想定し，高い応答精度くぉ実現する音声認識部と応答選択部の密結合システムを提案した．評価実験では，従来システムから 2.6% の改善が得られ，その有効性が示された．今後の課題は，汎用言語モデルの対話用言語モデルへの拡張などが挙げられる．

参考文献

[1] 福田敏則，吉見孔孝，南角吉彦，李晃伸，徳田恵一，“ユーザ生成型音声対 話コンテンツを用いた音声情報案内システム”，電子情報通信学会技術研究報告，SP2009-94, pp.207-212, Dec.2009.

[2] 加藤杏樹，南角吉彦，李晃伸，徳田恵一，“音声対話システムのためのキーワードの共起制約に基づくスポッティングア ルゴリズムの評価”，信学技報，vol.110, no.357, pp.25-30, Dec.2010.

[3] Y.Yoshimi, R.Kakitsuba, Y.Nankaku, A.Lee, and K.Tokuda, “Probabilistic Answer Selection Based on Conditional Random Fields for Spoken Dialog System,” Proc. of ICSLP,pp.215–218, 2008.

[4] 平野隆司，南角吉彦，李晃伸，徳田恵一，“双方探索に基づく N-gram に基づくキーワードからの文生成”，日本音響学会 2011 年春季研究発表会，2-P-40(b), Mar.2011.