Vol.2012-SLP-92 No.3
2012/7/19

登録キーワードと汎用言語モデルを用いた音声認識部・応答選択部の密結合に基づく統計的音声対話システム

平野隆司[†1] 加藤杏樹[†1,†2] 南角吉彦[†1]
李晃伸[†1] 徳田恵一[†1]

本論文では，ユーザが対話コンテンツを登録可能なユーザ生成型の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．ユーザ生成型のシステムでは，対話コンテンツは質問キーワードと対応する応答文の組として登録されるため，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，本論文では，登録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ガーベージによって密結合するシステムを提案する．評価実験において，十分な量の対話コーパスを学習した理想的なシステムにより得られる応答精度の上限を 91.8% としたとき，提案法では従来法であるキーワードに基づくシステムの 75.0% から 57.1% の誤り改善率が得られた．

A statistical spoken dialogue system based on tight integration of speech recognition and response selection using registered keywords and general language models

Takashi Hirano,[†1] Aki Kato,[†1,†2]
Yoshihiko Nankaku,[†1] Akinobu Lee[†1]
and Keiichi Tokuda [†1]

This paper describes an improvement of a user-generated spoken dialogue system. In the system, we assume users can freely register arbitrary dialogue content as a set of pairs of query keywords and a corresponding response sentence. Thus, at the dialog system, the speech recognition part and response selection part are usually connected by the keywords. However, the whole result of the speech recognition may include essential information for response selection, not only the keywords. Therefore, this paper proposes a spoken dialogue system in which the speech recognition part and response selection part are tightly connected by fully utilizing a general language model as garbage model and the registered keywords at each part. In the experiment, an ideal system in which all models are trained by a large amount of real dialogue corpus could achieve a response accuracy of 91.8%, whereas our proposed method achieved a 57.1% relative error reduction rate from the conventional keyword-based system that could achieve only 75.0% of response accuracy.

1. はじめに

近年，初見でもわかりやすい直感的な操作方法として音声インターフェースが注目を浴びている．そして，音声インターフェイスを利用したシステムに音声対話システムがある．

音声対話システムでは，大量のデータを収集し，学習した統計モデルに基づいて音声を認識したり，応答を選択する統計的手法の研究が盛んに行なわれている[1)2)]．その中でも，本研究では，一問一答形式の統計的音声対話システムを対象とする．統計的手法を用いた場合，システムのタスクに沿ったデータを用いてモデル学習を行うことで，高い応答精度が期待できる．しかし，実用システムに組み込む場合，モデル学習において大きな問題がある．

まず一つめは，統計学習には膨大な学習データを必要とする点である．一般的に，音声認識では大量の発話音声と膨大なテキストコーパスを必要とする．また，応答選択では質問文と応答文を組とする書き起こしたデータが必要である．さらに，自発的な発話に近いほど発話表現はユーザの個人性が大きくなる．そのため，ある対話タスクにおいてあらゆるユーザが発話しうる全ての発話をカバーするようなコーパスを構築することは多大な労力を要する．

二つめには，質問・応答の内容など，必要な統計情報がタスクに強く依存する点である．応答選択において，質問と応答を関連付けるためにはタスク特有の知識が必要不可欠となる．例えば，新規にシステム構築を考えた場合，タスク知識を個別に収集する必要があり，局所的なタスクであればあるほど，収集する量は増大する．そのとき，システム構築者は対

†1 名古屋工業大学大学院工学研究科
Graduate School of Engineering, Nagoya Institute of Technology
†2 現，ブラザー工業株式会社
Brother Industries, LTD.

話データを収集する必要があるが，考えうる全てのタスク知識を含む対話データを書き起こすには限界がある．また，各種イベントのお知らせなどの即時性の高いタスクを想定する場合，逐一，システム構築者が対話データを収集することは実用的ではない．

一方で，我々はユーザ生成型音声対話コンテンツの研究に取り組んできた[3)]．この研究では，ユーザ自身に質問文に含まれるキーワードとそれに対応する応答文を登録してもらうことで，対話コンテンツを収集し，提供している．この手法を用いることで，タスクをユーザが簡単に追加・記述できるようになり，より可搬性の高い音声対話システムが実現できる．しかし，タスク知識としてキーワードと応答文しか与えられていないため，ユーザとのロバストな音声対話を行うための統計知識が不足する．キーワードやタスクが増えるたびに依存した応答の統計情報やタスクごとの発話をカバーするようなコーパスを構築することは現実的ではない．

そこで，本研究では，ユーザ生成型音声対話コンテンツを想定し，登録キーワードと応答文の組のみが与えられる条件下で，より頑健で精度の高い音声対話システムの枠組みを検討する．具体的には，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を利用することで密に結合する．このとき，他に使用できる情報として汎用言語モデルを考え，汎用言語モデルに基づいてキーワード以外の部分を生成する．音声認識部ではディクテーションの探索中に動的にスポッティングを行うことで，キーワードを正確に抽出しつつ，文として出力する．応答選択部ではキーワード以外の部分を補間することで．学習用質問文を自動で生成する．これらの手法により，音声認識部と応答選択部の結合において，キーワードだけではなく，それ以外の部分でも対応付けられるので，より多くの有用な情報を応答選択に利用でき，応答性能の向上が期待できる．この手法によって，キーワードと応答文という少ないリソースを与えるだけで，高い応答精度が得られるシステムを構築できる．

以下，2 節で統計的音声対話システムの定式化による説明を行い，キーワードを用いた音声対話システムの紹介をする．3 節と 4 節では音声認識部と応答選択部の密結合に基づく統計的対話システムについて説明をする．

2. 一問一答形式の統計的音声対話システム

2.1 統計的音声対話システムの定式化

一問一答形式の質問応答における統計的手法では，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なる．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，質問発話に対して出力確率が最大となるような応答文 $\hat{A}$ は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，中間表現として認識した単語列 W を定義することで，以下の式のように置き換えられる．

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

一般的な統計的音声対話システムにおいて，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る．

式 (2) のシステムは，質問発話の音声信号から応答を選択する一般的な枠組みであり，それぞれの確率モデルに対して適切な学習がなされることが重要である．音声認識部 $P(W|O)$ のモデル学習では，大量の発話音声データとテキストコーパスを必要とする．また，応答選択部 $P(A|W)$ においては，発話文と応答文の組からなる対話データを大量に必要とする．しかしながら，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．このとき，対話データの収集コストと有用な情報の量はトレードオフの関係にある．

2.2 キーワードを用いた統計的音声対話システム

式 (1) の中間表現にキーワードを用いたシステムがある．この枠組みでは，応答選択の有用な情報になりうる発話単語をキーワードとして登録することで，応答精度の低下を抑制しつつ，対話データを収集する労力を削減できる．キーワードで対応付けた統計的音声対話システムには，最も簡単な実装方法であるディクテーションした単語列からキーワードのみを用いる手法や，キーワードスポッティングを用いた手法などが考えられる．

ディクテーションした単語列からキーワードのみを用いたシステムはキーワードを K とすると，以下の式で表せる．

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

式 (2) のシステムと式 (3) のシステムを比較すると，応答選択部 $P(A|K)$ では学習データの収集が簡略化できるが，キーワード抽出 $P(K|W)$ の前処理が必要となる．キーワード抽出 $P(K|W)$ には，あらかじめ登録した単語を正確に抽出できるテキストマッチングの手法などが考えられる．また，応答選択部 $P(A|K)$ ではキーワードのみを必要とするので，前段階の音声認識部 $P(W|O)$ においては，発話文全体の認識精度よりキーワードの認識精度が重要となる．

一方で，キーワードの認識精度を高めた手法にキーワードスポッティングがある．システ

ムは以下の式で表せる．

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

式 (4) 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，システム全体の応答性能の向上が期待できる．

3. 登録キーワードと汎用言語モデルを用いた統計的音声対話システム

本研究では，1 章で述べたとおり，ユーザ生成型音声対話コンテンツ[3] を想定しているので，キーワードとそれに対応する応答文のみが与えられると仮定する．その場合，応答選択部はキーワードのみで学習がなされるので，式 (1) の中間表現にはキーワードを用いる必要がある．しかし，質問発話にはキーワード以外の発話部分を表すガーベージが存在し，応答選択を補助する情報が含まれていると考えられ，応答精度の向上が期待できる．そこで，ガーベージを用いた音声対話システムの枠組みについて検討する．

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．

$$\hat{A} = \underset{A}{\operatorname{argmax}} \sum_{\{K,G\}} P(A|K,G)P(K,G|O) \quad (5)$$

ここで，$\{K,G\}$ は質問発話の書き起こし文（単語列）を表しており，W と同じような内容であるが，$\{K,G\}$ は文中に含まれるキーワード K が特定されている点で異なる．式 (5) の応答選択部 $P(A|K,G)$ と音声認識部 $P(K,G|O)$ のモデル化についてそれぞれ考える．

応答選択部 $P(A|K,G)$ は，モデルとしては式 (2) の $P(A|W)$ と同じ構造をしている．よって，その学習には応答文 A と対応する文 W が必要となる．しかし，ユーザ生成型の音声対話システムを考えた場合，学習には A と K のセットのみが与えられるため，なんらかの方法で 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ガーベージを補間し，文として自動生成する手法を提案する．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_w とると，W' は以下の式のように生成される．

$$\hat{W'} = \underset{W}{\operatorname{argmax}} P(W|K,\lambda_k,\lambda_w) \quad (6)$$

ここで，得られた W' を用いて応答選択部を学習した場合，応答選択部はキーワードの情報 λ_k に加え，一般的な言語情報 λ_w も用いているため，$P(A|W',\lambda_k,\lambda_w)$ と書くことができる．

次に音声認識部 $P(K,G|O)$ について考える．先に述べたように $\{K,G\}$ は文であるため，$W' = \{K,G\}$ と表すと音声認識部は $P(W'|O)$ と表現できる．これは，一般的な音声認識デコーダを表す．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，キーワードはガーベージよりも重要な情報を含むと考えられるが，一般的なデーコーダはキーワードとガーベージを区別していないため，音声信号から応答選択に必要な情報を抽出するという意味では最適とはいえない．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の仕組みが必要となる．

そのため，2.1 節で述べた式 (3) と式 (4) の音声認識部を応用し，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手法を検討する．この手法では，キーワード情報を用いてキーワードを抽出しつつ，ガーベージ部分は言語モデルにより補間される．このとき，言語モデルには汎用言語モデルを利用し，生成されたガーベージは除去しないで応答選択に活用する．このようなシステムでは，一般的な言語情報 λ_w に加え，キーワードの情報 λ_k も用いることから $P(W'|O,\lambda_k,\lambda_w)$ と記述できる．

上記のような考察を踏まえたうえで，もう一度，ガーベージを用いた対話システム (式 (5)) を書き直すと，

$$\hat{A} = \underset{A}{\operatorname{argmax}} \sum_{W'} P(A|W',\lambda_k,\lambda_w)P(W'|O,\lambda_k,\lambda_w) \text{ ただし，} W' = \{K,G\} \quad (7)$$

と書くことができる．このシステムは，一見，通常のディクテーションに基づくシステム (式 (2)) と同じに見えるが，キーワード情報 λ_k と一般的な言語情報 λ_w を応答選択部，音声認識部の双方に利用している点が異なる．すなわち，提案システム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．

4. 提案システム

4.1 全 体 構 成

ここでは，式 (7) に基づいて本研究で実際に構築したシステムの詳細について述べる．汎用言語モデルには，Web から収集したテキストから学習したタスクを限定しない 3-gram 言語モデル[4] を用いる．さらに，式 (7) では中間表現として全ての W' を考慮しているが，今回は $P(W'|O,\lambda_k,\lambda_w)$ が最大となる W' のみを使用する．提案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 $P(W'|O,\lambda_k,\lambda_w)$ が最大となる $W' = \{K,G\}$ を出力する．ここでは，登録キーワードと N-gram を用いてディクテーション中に動的に

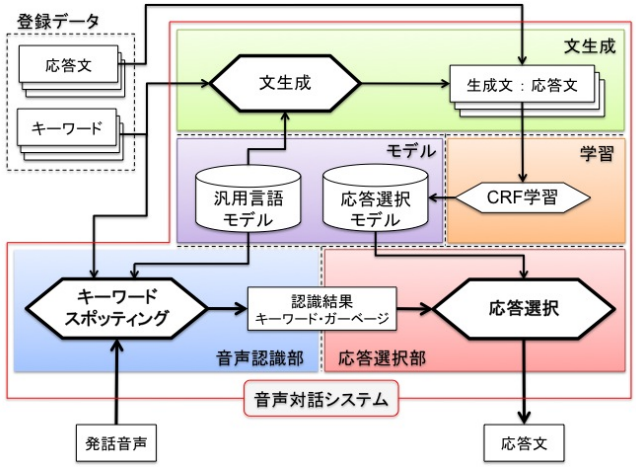

図 1 提案システムの全体構成
Fig. 1 An overview of the proposed system.

キーワードスポッティングを行う．このとき，N-gram はガーベージの言語モデルとして利用する．さらに，登録されるキーワードは複数で与えられることが想定されているので，キーワード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[5]．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 $P(A|W', \lambda_k, \lambda_w)$ 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．応答選択モデル $P(W'|O, \lambda_k, \lambda_w)$ は，条件付き確率場 (Conditional Random Fields; CRF)[6] で表現する．モデルの事前学習では，N-gram に基づいて登録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[7]．その後，生成された学習データベースのうち生成文のみ形態素解析を行い，応答選択モデル $P(W'|O, \lambda_k, \lambda_w)$ を学習する．

4.2 複数キーワードの共起制約に基づくスポッティング

音声認識部 $P(W'|O, \lambda_k, \lambda_w)$ では，ガーベージモデル（汎用言語モデル）を用いて探索中に動的にキーワード間制約を適用し複数キーワードのスポッティングを行う[5]．この手法では，ガーベージとキーワードの任意の繰り返しを許す連接モデルで表現し，キーワード間制約を探索中に動的に与える．探索時のキーワード間の出現制約の与え方として，キーワードセット内の出現順を考慮する順列制約を用いる．このような制約を与えることで，定義にないキーワードセットの出現を抑えられる．

4.3 自動文生成に基づく CRF を用いた応答選択

応答選択部 $P(A|W', \lambda_k, \lambda_w)$ では，N-gram に基づく登録キーワードからの自動文生成手法[7] と，CRF に基づいた応答選択手法を用いる[8]．

まず，自動文生成手法では，与えられた N-gram 言語モデルに基づいて任意のキーワードを含む最も単語の連接確率が高い文を生成する．具体的には，キーワードおよび文開始記号 $< S >$・文終了記号 $< /S >$ に挟まれた区間ごとに単語列をそれぞれ探索する．

その後，各区間ごとに得られた単語列とキーワードを組み合わせて，リスコアリングを行い，最尤の文を効率よく生成する．このとき，単語列の探索時には，探索範囲が広くなりすぎるので，ビーム探索や双方向探索などを行う．最終的に，生成した文のうち尤度の上位 N 文を選択し，対応する応答文との組を CRF 用の学習データとして使用する．

次に CRF に基づく応答選択では，入力である発話文と出力である応答文の間にある特徴を表す素性関数とその重みを用いてモデルを表現し，応答文選択を行う．

CRF の学習では，生成された単語列の各単語に対して応答文の ID を割りつけて学習する．応答選択時は，単語系列に含まれるそれぞれの単語に対して応答文が出力されるが，応答文が全て一致する候補の中で確率が最大のものを選択し，出力する．

また，この応答選択手法においては，入力単位が文もしくはキーワードの場合に，学習データは入力単位に合わせて，文もしくはキーワードと応答文の組を使用する．

5. 評価実験

5.1 実験条件

一問一答形式の音声情報案内システム「たけまるくん」[9] のタスクにおいて，タスク知識が十分である場合と少ない場合での対話システムの構築を想定し，応答性能を比較する．

使用した音声データは，「たけまるくん」で録音したものである．対話データは，人手で録音データから書き起こし，正解の応答文と対応付けたものである．

タスク知識である登録キーワードは対話データの質問文から，人手でキーワードを抽出した．汎用言語モデルには，World Wide Web から収集したテキストコーパスから学習した N-gram 言語モデルを使用した．また，比較のため，タスク言語モデルであるたけまるくん大人用言語モデルも用意した．キーワードスポッティングは Julius に実装し，応答選択には CRF++を使用した．

表 1 学習データセットとテストデータセット
Table 1 Training and testing data sets.

	学習データセット	テストデータセット
データ数	8543	527
キーワードセット種類	406	126
キーワード種類	320	114
キーワード延べ	13779	803
応答種類	151	72

5.2 実験用データベース

「たけまるくん」の対話データの書き起こしからキーワードと応答文データベースを作成した．「たけまるくん」のユーザ発話の書き起こし文を元に手作業でキーワードを抜き出した．書き起こしデータは，「たけまるくん」のユーザ発話のうちの有効発話の中から 2002 年 11 月から 2004 年 10 月までを学習セットとして，2003 年 8 月分をテストセットとして使用した．

実験に先立ち，データの整備を行なった．まず，キーワードの抽出時には，設定したキーワードが含まれていない場合は無効発話として除外した．さらに，発話文の書き起こしには，同一単語の漢字表記とひらがな表記の両方が含まれていたため，統一した．また，同一意図の発話文に対して異なる応答文が存在していたので，頻度が少ない方を削除した．

そして，「こんにちは」などの単単語発話に関しては，キーワードのみで成り立ち，それ以外のガーベージ部分が不要であると考えられる．そのため，本実験では対象外として除外した．キーワード抽出後の学習とテストのデータセットの詳細を表 1 に示す．

5.3 統計的音声対話システムの比較

提案する密結合システムの有効性を確かめるため，応答正答率を求めた．比較したシステムは以下の 3 つである．

- 「理想的なシステム」 タスクに沿った質問発話文と応答文の書き起こしデータが収集できる理想的な場合を想定し，構築したシステム．タスク知識が十分にあり，応答性能の上限に相当するシステムである．システム構成は，式 (2) に基づいてディクテーションと質問発話文と応答文を対応付けた CRF モデル $P(A|W)$ を用いた．言語モデルには，タスク言語モデルを使用した．
- 「キーワードシステム」 ユーザ生成型音声対話コンテンツを想定し，少ないタスク知識で構築した従来法にあたるシステム．「理想的なシステム」と比べて登録キーワードしか与えられない．システムは式 (3) に基づいて構築した． また，ディクテーションに用いる辞書にキーワードの情報を追加した．
- 「密結合システム」 ユーザ生成型音声対話コンテンツを想定し，音声認識部と応答選択部の密結合を意識した提案法にあたるシステム．

表 2 統計的音声対話システムの比較
Table 2 Comparison of statistical spoken dialogue systems.

統計的音声対話システム	理想的なシステム	キーワードシステム	密結合システム
応答正答率 (%)	91.8	75.0	82.2

構築したそれぞれのシステムで応答選択を行い，算出した応答正答率を表 2 に示す．

「理想的なシステム」と「キーワードシステム」を比較すると，学習データである質問発話文をキーワードにすることで，収集コストを抑えたために，応答正答率が 16.8% 低下した．しかし，「理想的なシステム」の応答正答率を上限としたとき，提案する「密結合システム」では「キーワードシステム」から 50.7% の誤り改善率が得られ，半分近く応答正答率が回復した．

5.4 音声認識部の評価

提案システムの音声認識部がシステム全体の応答性能に与える影響について詳細に評価する．応答選択部は，5.3 節とは異なり，文生成手法に基づく応答選択 $P(A|W')$ に統一する．比較する認識手法は以下の 3 通りである．

- 「ディクテーション」 ディクテーションによる音声認識．汎用言語モデルだけを利用しているので，応答選択部との結合力は弱い．
- 「キーワード追加辞書を用いたディクテーション」辞書にキーワードを追加したディクテーションによる従来法にあたる音声認識．「ディクテーション」と比べてキーワードの情報を用いている．
- 「キーワードスポッティング」 複数キーワードの共起制約に基づくスポッティングによる提案法にあたる音声認識．他の認識手法と異なり，キーワード及びセット単位での抽出性能を高めている．キーワードの情報を最大限利用することで，応答選択部との結合力は最も強くなる．

評価指標は，Precision と Recall, F 値[5] のキーワード正解率と，応答正答率を用いる．それぞれの認識手法を用いたときのキーワード正解率および応答正答率を表 3 にまとめて示す．

結果より，3 手法の中でシステム全体の密結合を意識した「スポッティング」が最も高い応答性能が得られた．しかし，「スポッティング」の F 値は「ディクテーション」よりも低い値になっている．さらに，Precision においても「ディクテーション」より低く，キーワー

表 3 音声認識手法の比較
Table 3 Comparison of speech recognition methods.

音声認識手法	応答正答率 (%)	Persicion	Recall	F 値
ディクテーション	72.7	0.8467	0.7497	0.7952
キーワード追加辞書を用いたディクテーション	79.1	0.8558	0.7908	0.8220
キーワードスポッティング	82.2	0.7791	0.7995	0.7892

表 4 応答選択手法の比較
Table 4 Comparison of responce selection methods.

モデル	理想的な CRF	キーワード CRF	生成文 CRF
応答正答率 (%)	86.5	78.0	82.2

ドの抽出結果の中に正解キーワードの数がより少なかったことを表している．一方，Recall を見ると，「スポッティング」が最も高い値を示しており，キーワードを検出できた総数が多かったことを表している．これらの結果から，「スポッティング」はキーワードの誤抽出の数は増加したが，抽出キーワードの数を増加させることで，システム全体の応答性能を高める手法であるといえる．

5.5 応答選択部の評価

次に，提案する応答選択部が応答性能に与える影響を評価した．具体的にはモデル構造が異なるを 3 通りの CRF を用意し，応答選択を行ことで，文生成手法の有効性を確かめた．使用するモデルは以下の 3 通りである．

- 「理想的な CRF」 質問発話文と応答文の組を収集し，学習した CRF $P(A|W)$ ．十分な量のタスク知識を含んでいる．
- 「キーワード CRF」 ユーザ生成型音声対話コンテンツを想定し，登録キーワードと応答文の組から学習した従来法にあたる CRF$P(A|K)$．「理想的な CRF」からキーワードを抽出し，作成した．「理想的な CRF」よりもタスク知識は少ない．
- 「生成文 CRF」 生成文と応答文の組から学習した CRF$P(A|W')$．「キーワード CRF」の登録キーワードから質問発話文を自動生成した．「キーワード CRF」に比べてタスク知識が補強される．

音声認識部では登録したキーワードを抽出するため，5.4 節の「キーワードスポッティング」を用いる．それぞれの CRF を用いて応答選択を行った結果を表 4 に示す．

タスク知識が少ない「キーワード CRF」は「理想的な CRF」に比べると応答正答率が 8.5% 低下した．しかし，「理想的な CRF」の応答性脳を上限としたとき，「生成文 CRF」では「キーワード CRF」から 50.6% の誤り改善率が得られた．

6. むすび

本稿では，ユーザ生成型の音声対話システムを想定し，音声認識部と応答選択部の密結合による高い応答精度を実現するシステムを提案した．評価実験では，提案システムでは従来システムから 57.1% の誤り改善率が得られ，その有効性が示された．今後の課題は，実システムへの導入などがある．

謝辞 本研究の一部は科研費 (21300066) の助成を受けたものである．

参考文献

1) 堀智織，翠輝久，大竹清敬，柏岡秀紀，中村哲：統計的対話モデルを用いた WFST に基づく音声対話システム，電子情報通信学会技術研究報告.SP，音声 108(422), pp.25-30, (Jan. 2009).
2) Williams J.D. and Young S.J. : Partially observable Markov decision processes for spoken dialog systems, *In proc. Computer Speech and Languag 21 (2)*, 231-422, (2009).
3) 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対話システム，信学技報，vol.109, no.356, pp.207-212, (Dec. 2009).
4) 北研二：確率的言語モデル，東京大学出版会，(1999).
5) 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).
6) Lafferty J., McCallum A. and Pereira F. : Conditional random fields : Probabilistic models for segmenting and labeling sequence data, *In proc. of ICML*, pp282-289, (2001).
7) 平野隆司，南角吉彦，李晃伸，徳田恵一：双方向探索に基づく N-gram を用いたキーワードからの文生成，日本音響学会講演論文集，vol.I, 2-P-40(b), pp.211-212, (Mar. 2011).
8) Yoshitaka Yoshimi, Ryota Kakitsuba, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda : Probabilistic Answer Selection Based on Conditional Random Fields for Spoken Dialog System, *In proc. Interspeech 2008*, pp.215-218, (Sep. 2008).
9) Ryuichi Nisimura, Yohei Nishihara, Ryosuke Tsurumi, Akinobu Lee, Hiroshi Saruwatari, Kiyohiro Shikano : Takemaru-kun : Speech-Oriented Information System for Real World Research Platform, *In proc. First International Workshop on Language Understanding and Agents for Real World Interaction*, pp.70-78, (2003).