社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－SLP－74（46）
2008 / 12 / 10

質問応答データベースの自動作成に基づく音声対話システムの評価

森本 高弘† 伊藤 仁† 鈴木 基之†† 伊藤 彰則† 牧野 正三†

† 東北大学大学院工学研究科
〒980-857 宮城県仙台市青葉区荒巻字青葉 6-6-05
†† 徳島大学大学院ソシオテクノサイエンス研究部
〒770-8506 徳島県徳島市南常三島町 2-1
E-mail: †{takahiro,itojin,aito,makino}@makino.ecei.tohoku.ac.jp, ††moto@m.ieice.org

あらまし 一問一答形式の音声対話システムにおいて質問応答データベースを用いた用例ベースの応答生成は様々な発話に頑強だが，新たなシステム設計はコストが大きい．本研究ではコスト削減のため，用例の種類ごとに用意されたテンプレートを用いて質問応答データベースを自動作成する方法について検討，評価を行う．結果，人手で作成した質問応答データベースを用いた場合と自動作成した質問応答データベースを用いた場合でとほぼ同様の性能を示すができた．また，F 値を用いたスコアリング方法を提案し，従来法を用いた場合よりも応答正解率が改善した．
キーワード 音声対話システム，Q&A データベース，自動作成，F 値

Estimation of Spoken Dialog System using Automatically-generated question-and-answer database

Takahiro MORIMOTO†, Masashi ITO†, Motoyuki SUZUKI††, Akinori ITO†, and Shozo MAKINO†

† Graduate school of engineering, Tohoku University,
6-6-5, Aoba, Aramaki-aza, Aoba-ku, Sendai, Miyagi, 980-8579, Japan
†† Institute of Technology and Science, The University of Tokushima,
2-1, Minamijosanjima-cho, Tokushima, Tokushima, 770-8506, Japan
E-mail: †{takahiro,itojin,aito,makino}@makino.ecei.tohoku.ac.jp, ††moto@m.ieice.org

Abstract A question-and-answer style spoken dialog system based on example-based answer generation is known to be robust against variation of user utterances. However, it is costly to create QA database for a new task. In this paper, we proposed a method to reduce cost of preparing the database by generating the database automatically from templates. As a result, we obtained almost same performance using the automatically generated QA database compared with the manually prepared database. In addition, we propose a new scoring method to choose an answer based on F-measure,which improved the accuracy of answer selection.
Key words Spoken dialog system, Q&A database, automatic database generation, F-measure

1. はじめに

近年，音声認識技術が発達し，音声対話システムが実現している．音声対話システムとは，ユーザの自然発話を認識し，その発話に対して音声によって応答を返すシステムである．扱うタスクがごく小規模である場合，認識に必要な単語は限られているため，認識文法を用いることは有効である．しかし，大規模なタスクの場合，新たにデータを追加していくと認識文法が破綻する可能性がある．このため，大規模タスクにおいては認識文法を用いずに新たなデータを追加できるシステムの方が適していると考える．「たけまるくん」[1] は認識文法を用いずに音声対話を実現している．このシステムでは，あらかじめユーザが行うであろう質問文（用例テキスト）とその答（応答候補文）の対のデータベース（質問応答データベース）が用意されている．用例テキストを用いた質問応答では，アドホックなデータの追加ができるため，柔軟性が高く，様々な発話に対応できるため，大規模タスクに向いていると考えられる．

しかし，用例を用いた音声対話システムを新たに開発する場

合，あらかじめ質問応答データベースを作成しておき，フィールドテストなどを行い様々な発話を獲得して，質問応答データベースを更新していくことになる．ユーザは同じ内容の発話でも様々な言い換えを行うため，非常に多くのデータベースが必要になる．システムの性能向上には質問応答データベースが最も重要であるとの報告 [2] もあり，データベースの整備は非常に重要である．ユーザがどのような発話を行うかは予想しきれないため，データベースを随時更新していく必要があるが，運営の段階であらかじめ予想ができる発話もあると考える．そこで，本稿ではシステム運用にあたってあらかじめ用意しておく質問応答データベースの作成にかかるコストの削減を目的に，テンプレートを用いてデータベースを自動作成する方法 [3] について述べる．また，一般のコーパスから作成した言語モデルではタスクの表現には対応できないため，発話の認識誤りが多く起こる．認識誤りを減らすため，自動作成したデータベースを用いたタスク適応について検討する．最後にシステムの性能の向上のため，F 値を用いたスコアリング方法を提案する．

2. 用例テキストを用いる音声対話

用例テキストを用いる音声対話システムの概要を図 1 に示す．ユーザがシステムに発話を行うと，システムは大語彙連続音声認識によって発話を認識し，その発話の認識文と用例テキストの間でマッチングを行い，各用例テキストにスコアリングを行う．スコアが最も高い用例テキストを認識文に最も似ている用例テキストと判断し，その用例テキストに対応している応答候補文を応答としてユーザに返す．

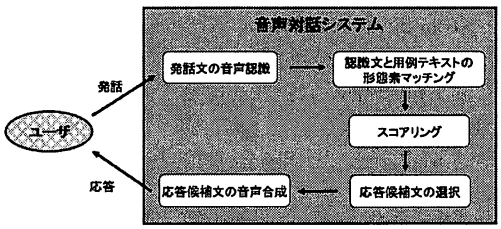

図 1 システムの概要

システムにはあらかじめ用例テキストと応答候補文が用意されている．一つの応答候補文に対して複数の用例テキストが用意されている．これは，ユーザは同じ内容でも異なる表現を用いるためであり，複数の用例テキストを用意することにより，ユーザの様々な発話に対応できる．

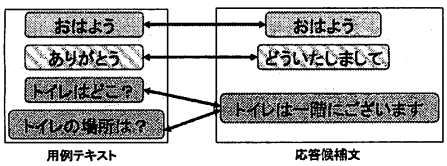

図 2 用例テキストと応答候補文

ユーザに返す応答候補文を選択するために，発話の認識文と用例テキストとの間で自立語形態素のマッチングを行う．自立語形態素とは，名詞や代名詞といった，単独で文節を構成でき，文章において意味を持つ最小単位である．マッチした自立語形態素数に応じて用例テキストにスコアリングを行う．全用例テキストのスコアリングをした後，各応答候補文のスコアを決定する．同じ応答候補文に対応する用例テキストの中で，最も高い用例テキストのスコアをその応答候補文のスコアとする．全応答候補文のスコアリングをし，最もスコアが高かった応答候補文をユーザの発話に対してふさわしい応答であるとして選択し，この応答候補文をユーザに応答として返す．

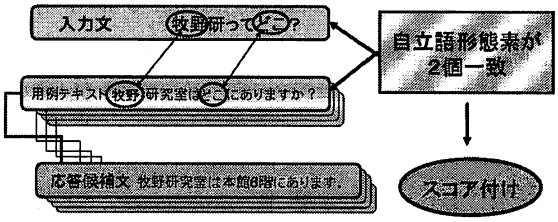

図 3 スコアリング

3. 質問応答データベースの自動作成

3.1 質問応答データベースについて

従来の質問応答データベース作成では用例テキストと応答候補文を一つ一つ人手で記述しており，非常に時間と手間がかかる．用例テキストは「キーワード」と「特定の言い回し」の組合せでできているものが多数ある．「牧野研究室はどこですか」という用例テキストは「牧野研究室」というキーワードと「はどこですか」という場所を聞く質問の言い回しに分けることができる．この用例テキストの言い換えを作るときには，キーワードのみを言い換える場合 (「研究室」を「研」に) と，言い回しのみを変える場合 (「はどこですか」を「に行きたいのですが」に) と，その両方を変える場合が考えられ，データベースの整備は非常に大変となる．人手で作成する場合，「キーワードの言い換えの数×言い回しの数」の用例テキストを全て書きおこさなくてはならない．

本稿では用例テンプレート，変換ルール，言い換えルールを用意しておき，データベース記述表に必要な情報を記載することにより用例テキストと応答候補文が自動的に作成できる方法を提案する．「牧野研はどこですか」を言い換えた用例テキストを作成するために必要な記述は，「言い換えルールの数＋用例テンプレートの数」個のデータ，キーワードおよび作成する用例の種類を記述したデータベース記述表である．これらの記述から用例テキストが自動的に作成されるため，データベース作成にかかるコストを大幅に減らすことができる．

3.2 データベース自動作成の流れ

予想される入力よりいくつかの分類を定め，その分類に応じた言い回しを記述したテンプレート（用例テンプレート）と変換ルールと言い換えルールを用意しておく．データベース記述表にはキーワード等の必要な情報を記載する．データベース記

述表よりキーワードとその答を組み合わせて，応答候補文が作成される．キーワードが変換ルール，言い換えルールに当てはまるか判断し，当てはまる場合は変換，言い換えを行う．その後，キーワードと用例テンプレートを組み合わせて，用例テキストが作成される．

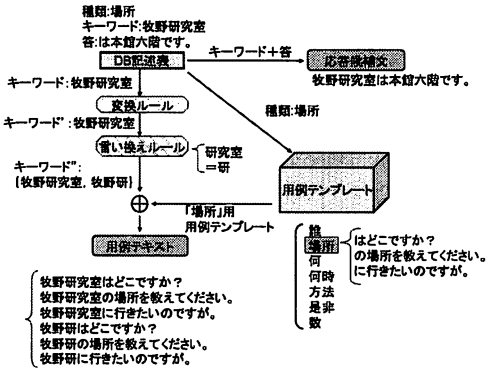

図 4　データベースの自動作成

3.3　データベース記述表

質問応答データベースを作成する際に必要な情報を記載したものを，データベース記述表と呼ぶ．表には以下の項目を記載する．

- 質問の種類（以下「種類」）
- キーワード（種類が「挨拶」以外）
- キーワードに対する答（以下「答」，種類が「挨拶」以外）
- 用例テキスト（種類が「挨拶」のみ）
- 応答候補文（種類が「挨拶」のみ）

種類を「誰」「場所」「何」「何時」「方法」「是非」「数」「挨拶」の 8 分類に分類する．「挨拶」に限り，キーワードと答は用意せず，用例テキストと応答候補文を記載する．キーワードとは応答候補文において主語になる部分，答はキーワードに続く文と定義する．

表 1　データベース記述表

種類	キーワード	答	用例テキスト	応答候補文
挨拶	-	-	ありがとう	どういたしまして
場所	ここ	は一号館です。	-	-
誰	教授	は牧野正三先生です。	-	-

3.4　用例テンプレート

先述した種類別に，キーワードに続く文を記載したものを用例テンプレートと呼ぶ（「挨拶」には用意されていない）．キーワードと組み合わせる事により，質問文が構成される．

表 2　用例テンプレート

誰	場所	方法
は誰ですか？	はどこですか？	はどうやりますか？
はどなたですか？	はどこにありますか？	のやり方を教えて下さい？
の名前は何？	の場所を教えて。	はどうすればいいのですか？

3.5　変換ルール

キーワードは，答の主語であるが，必ずしも質問の主語とは一致しない．「私の名前は QuA（システムの名前）です.」に予想される質問文は「あなたは誰ですか？」などであり，主語が「私」と「あなた」と異なる．このように答と質問文の主語が異なる場合のために変換ルールを導入し，主語のずれを無くす．

3.6　言い換えルール

意味が等価であると考えられる別の表現を言い換えと呼ぶ．キーワードが複数の言い換えを想定できる場合，キーワードを言い換えて用例テキストを作成する．

表 3　言い換えルール

研究室 = 研
トイレ = お手洗い = 化粧室
一号館 = 本館 = ここ

3.7　評価実験

3.7.1　想定するタスク

システムの扱うタスクは，音声情報案内システムを東北大学工学部電子・応物系一号館に設置すると想定して定めた 6 つのカテゴリのタスク (表 4) である．以下，このタスクで評価実験を行う．

表 4　タスクの詳細

カテゴリ	内容	想定する内容の例
greeting	挨拶	こんにちは
system	システムに関する情報	あなたの名前は？
facility	電子・応物系施設案内	2 号館はどこ？
laboratory	電子・応物系研究室場所案内	牧野研に行きたい
makino	牧野研究室に関する情報	牧野研は何人いる？
time	現在の時刻	今何時？

3.7.2　応答正解率

通常の応答正解率では質問応答を行い，入力の数のうち相応しい応答が出力された割合で表されるが，複数の応答候補文のスコアが最高スコアとなった場合，その中から出力する応答候補文をランダムに選ぶため，各手法を比較するのには不都合である．そこで性能の評価の指標として，平均応答正解率を求める．平均応答正解率 (以下，応答正解率) は式 (1) で表される．

$$\text{平均応答正解率} = \frac{\sum_i x_i}{\text{入力の数}} \tag{1}$$

$$x_i = \frac{\alpha}{\text{入力 } i \text{ に対し最高スコアを持つ応答候補文数}} \tag{2}$$

$$\alpha = \begin{cases} 1 & \text{最高スコア群に正解が含まれる場合} \\ 0 & \text{それ以外} \end{cases} \tag{3}$$

3.7.3 自動作成データベースの評価

人手で作成したデータベース（人手作成 DB），図 4 の流れで作成したデータベース（自動作成 DB）の応答正解率を求め，タスク別に比較を行う．音声入力は行わず，テキスト 225 文 (表 5) を入力として用いた．

質問応答データベースの詳細は表 6 に，データベースを自動作成するために用意した用例テンプレートなどの詳細を表 7 に記載する．データベース記述表の数は「種類」「キーワード」「答」を一組と，「変換ルール」，「言い換えルール」はルール一つを一組と数えている．実験結果は図 5 となった．

表 5　入力文の詳細

カテゴリ	入力 (文)
greeting	25
system	48
facility	42
laboratory	42
makino	45
time	23
合計	225

表 6　質問応答データベースの詳細

カテゴリ	人手作成用例 (文)	自動作成用例 (文)	応答 (文)
greeting	26	26	12
system	61	98	14
facility	253	766	41
laboratory	172	946	43
makino	87	363	19
time	11	10	1
合計	528	2209	130

表 7　データベース自動作成詳細

データベース記述表	119 組
言い換えルール	11 組
変換ルール	1 組
用例テンプレート	43 個

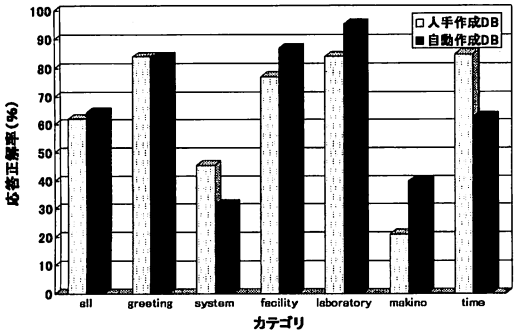

図 5　応答正解率

人手で作成した場合よりも全体の正解率は若干向上した．カテゴリ別に見ると，「facility」「laboratory」「makino」で応答正解率が増加した．これは，自動作成により，キーワードと言い回しの全ての組合せの用例テキストが用意できたためである．一方「system」「time」では応答正解率は減少した．これは，人手では単語単位や構造単位での言い換えを行い，用例テキストを作成できる一方で，本手法の現在の枠組みでは言い換えは単語単位に留まり，また「キーワード＋用例テンプレート」の形以外の用例テキストが作れないため，適切な用例テキストが作成できないためである．例えば，「今の時間は何時ですか」の言い換えである「何時ですか」という用例テキストは作成できない．「キーワード＋用例テンプレート」の形以外の用例テキストも自動作成できる枠組みが必要である．

4. 用例テキストを用いたタスク適応

音声認識のために様々な言語モデルが配布されているが，音声対話システムの想定する特定のタスクに対応するのは難しい．そこで，一般的な内容のコーパスと自動作成した用例テキストから言語モデルを作成する．また，一般的なコーパスと比較し，自動作成した用例テキストのデータは非常に少ないため，単純にコーパスに用例テキストのデータを加えただけでは用例テキストを混ぜる影響は小さいと考えられる．そこで，n-gram カウント混合 [5] により，タスクに適応した言語モデルの作成を目指す．

4.1 評価実験

実験に用いたデータと言語モデル作成条件を表 8 に示す．表 5 のテキスト文の読み上げ音声 225 発話を音声認識し，単語正解率 (Corr.) と単語正解精度 (Acc.) を求める．また，音声認識結果を入力して質問応答を行い，応答正解率を求める．本実験では一般的な内容のコーパスとして CSJ [6] を選択した．まず，CSJ 講演のみ，用例テキストのみ，CSJ 講演と用例テキストの混合 (混合比 1:1) から学習した 3 つの言語モデル (それぞれ CSJ-LM，用例-LM，CSJ+用例-LM) を評価する．次に CSJ の講演に用例テキストの混合比を変化させて混ぜ合わせて学習した言語モデルを評価する．

用いた自動作成質問応答データベースは表 6 と同様のもの，音声認識の条件は表 9，質問応答の条件は表 10 である．

表 8　言語モデル作成条件

一般的なコーパス	CSJ 学会・模擬，対話 2580 講演書き起こし文，22MB
用例テキスト	自動作成用例データ 2209 文，65KB
言語モデル語彙	頻度 1 回以上の単語

表 9　音声認識条件

入力音声	男性話者 10 名の読み上げ音声 225 発話
言語モデル	3-gram，good-turing 平滑化
音響モデル	CSJ 性別非依存モデル
認識エンジン	Julius-3.5.3

表 10　質問応答条件

入力	表 9 の条件で認識した 225 発話
用例テキスト	2209 文
応答候補文	130 文

4.1.1　各言語モデルとの比較

CSJ-LM，用例-LM，CSJ+用例-LM を用いた際の音声認識率とその音声認識結果を用いた際の全体の応答正解率を表 11，カテゴリ別の応答正解率を図 6 に示す．

表 11　各モデルの音声認識率とその応答正解率

言語モデル	Corr.[%]	Acc.[%]	全体の応答正解率 [%]
CSJ-LM	56.52	50.96	32.42
用例-LM	48.04	44.55	37.33
CSJ+用例-LM	64.22	61.87	51.04

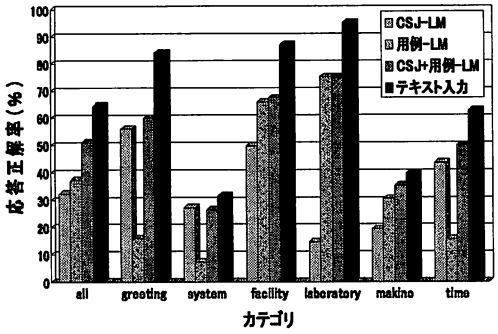

図 6　各言語モデルを用いた際の応答正解率

CSJ と用例を混ぜ合わせて学習することで，音声認識率，応答正解率ともに改善した．CSJ-LM を使った場合，用例-LM と比べて音声認識率が高いが，応答正解率が低い．その理由は，CSJ のコーパスは非常に多くの表現を含んでいるため様々な表現に対応できるが，タスク特有の表現を含んでいないためキーワードを認識できず，正しい応答候補文を選べない場合が多いためである．

用例-LM では用例テキストの総単語数が少なく，また各単語の出現頻度のばらつきが大きい．このため，他のカテゴリと比較して用例テキストが少ないカテゴリ（「greeting」「system」「time」）の単語は用例テキスト中の出現頻度が低いため，出現頻度の高い単語への誤りが多くなり，不正解となる（間違いの例：「今晩は」->「本館は」）．両者の言語モデルには問題があるが，CSJ と用例を混ぜることにより，CSJ，用例単独で学習した言語モデルよりも応答正解率は改善した．これは総単語数が多くなった結果，出現頻度のばらつきの影響が減り，タスク特有の表現を含むコーパスを学習することでタスクのキーワードを認識できるようになったためである．一般のコーパスと自動作成した用例テキストを混ぜ合わせることにより，それぞれのコーパスのみから作成した言語モデルの弱点を改善するができた．

4.1.2　N-gram カウント混合による言語モデルの評価

CSJ に対して重みを付けて用例テキストを混ぜ合わせて学習した言語モデルを用いた際の応答正解率を図 7 に示す．

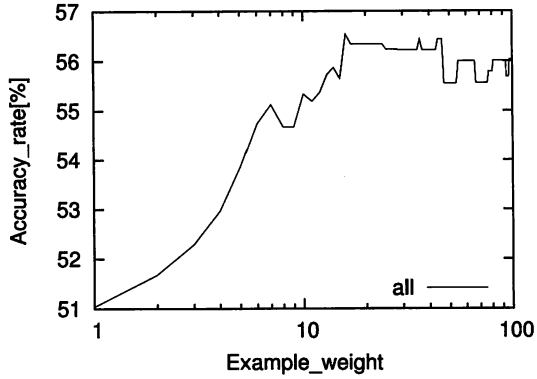

図 7　重み付き学習した際の応答正解率

重みを付けて言語モデルを作成することにより応答正解率が改善した．重みが 15 付近で応答正解率は最大となり，重み 1 の時と比較し，5 ポイント程度の改善が見られた．重みを増すことで用例テキストに現れる単語列の出現頻度が増し，よりタスクに適応した言語モデルの作成ができたと考えられる．

5.　F 値によるスコアリングの検討

従来法ではユーザの入力に対して各用例テキストのスコアを式 (4) で求め，このスコアから応答候補文のスコアを式 (5) で求めている．ここで s は入力文，e は用例テキスト，$C(x)$ は文 x に含まれる自立語形態素の集合，$N(w,x)$ は文 x に含まれる単語 w の個数，$N(x)$ は文 x の自立語形態素数，E_i は i 番目の応答候補文に対応する用例テキストの集合である．

$$r(s,e) = \frac{\sum_{w \in C(s)} \min\{N(w,s), N(w,e)\}}{N(e)} \tag{4}$$

$$SP_i(s) = \max_{e \in E_i} r(s,e) \tag{5}$$

しかしこの方法は想定するタスクによっては問題がある．例えば，タスクとしてトイレを聞く質問を想定しているとする．ユーザが「トイレはどこですか」といった場合，男子トイレか女子トイレどちらかを尋ねる質問かをシステムは区別がつかないので，両方のトイレの場所を教えるべきである．ユーザが「男子トイレはどこですか」と尋ねた場合，女子トイレの場所の情報はいらず，男子トイレのみの場所を教えるべきである．このような状況を想定すると「トイレはどこですか」「男子トイレはどこですか」という別々の用例テキストを用意することになる．しかし，「男子トイレはどこ」といった入力があった場合，従来法では両方の用例テキストのスコアが同じになってしまうため正しい応答を返せない．これは，入力文の自立語形態素を考慮していないことが原因である．そこで F 値を用いたスコアリング手法を提案する．従来法でのスコア $r(s,e)$ は，用例 e

に対する入力 s の再現率と見なすことができる。これに対して s の適合率を式 (6) で求める.

$$p(s,e) = \frac{\sum_{w \in C(s)} \min\{N(w,s), N(w,e)\}}{N(s)} \quad (6)$$

最終的に，再現率と適合率から F 値を求め，これをスコアとする.

$$F(s,e) = \frac{2p(s,e)r(s,e)}{p(s,e) + r(s,e)} \quad (7)$$

$$SF_i(s) = \max_{e \in E_i} F(s,e) \quad (8)$$

5.1 評価実験

従来法と提案法でスコアリングを行い，それぞれ応答正解率を算出した．実験にはテキスト入力と音声入力の両方を行う．音声入力では N-gram カウント混合を用いた言語モデルの評価において応答正解率が最も最高となった，CSJ に対して用例テキストの重みを 16 として混ぜ合わせて学習した言語モデルを用いた．他の実験条件は表 6，9，10 と同様である．音声認識結果を表 12 に，実験結果を図 8 に示す.

表 12　N-gram カウント混合言語モデルを用いた際の音声認識率

Corr.[%]	Acc.[%]
68.50	66.22

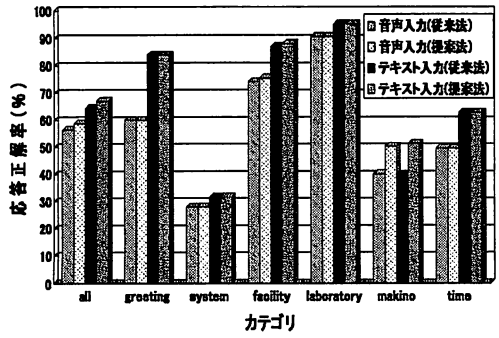

図 8　F 値をスコアリングに用いた際の応答正解率

応答正解率は約 2 ポイントの改善がみられた．これは，従来法においては自立語形態素を多く含む長い入力文はマッチングしやすくなるため，用例テキストのスコアが高くなる傾向がある．特に自立語形態素が少ない短い用例テキストはスコアが高くなるため，短い用例テキストが用意されている応答候補文を選ぶ誤りが多く発生してしまう．しかし提案法においては，入力文の自立語形態素数を考慮するため，入力文が長いほどスコアが高くなる傾向を抑えることができ，短い用例テキストが用意されている応答候補文を選ぶ誤りが減ったためである.

6. おわりに

本稿では，今まで人手で用意していた質問応答データベースを質問の種類毎に自動作成することを提案し，人手で作成した質問応答データベースを用いた場合とほぼ同様の性能を得ることができた．そして一般的な内容のコーパスに対して，自動作成した用例テキストに重みをつけて n-gram 混合することにより，タスクに適応した言語モデルを作成することができた．また，F 値を用いたスコアリングにより，入力文を考慮しないことによる問題が解決され，応答生成の性能が改善した．今後は本手法が他のタスクでも有効かどうかを確かめるとともにデータベース自動作成の枠組の確立をめざす.

文　献

[1] 西村 他，"実環境研究プラットホームとしての音声情報案内システムの運用"，信学論，Vol.J87-D-II，No.3，pp.789-798，2004

[2] トピアス 他，"「たけまるくん」実環境音声案内システムのデータベース整備と「キタちゃん」へのポータビリティーの検討"，情処研報，SLP-64-31，pp.173-178，2006.

[3] 森本 他，"質問種類別テンプレートを用いた質問応答データベース自動作成"，音講論，pp.107-108，2008-3.

[4] 森本 他，"自動作成用例データを用いたタスク適応言語モデルの検討"，音講論，pp51-52，2008-9.

[5] 伊藤，好田，"N-gram 出現回数の混合によるタスク適応の性能解析"，信学論，Vol.J83-D-II，No.11，pp.2418-2427，2000

[6] 独立行政法人国立国語研究所，日本語話し言葉コーパス，2004.