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

2003－NL－155 (16)
2003－SLP－46 (5)
2003／5／27

ニュース音声を対象とした音声質問応答システムの試作

西崎 博光† 中川 聖一††

† 山梨大学大学院医学工学総合研究部
〒 400-8511 山梨県甲府市武田 4-3-11
†† 豊橋技術科学大学情報工学系
〒 441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

E-mail: †nisizaki@ccn.yamanashi.ac.jp, ††nakagawa@slp.ics.tut.ac.jp

あらまし 本稿では，音声文書検索の発展として，音声入力を用いた質問応答システムについて述べる．現在の情報検索システムの多くは，ユーザが入力したクエリーに対する回答として文書全体を返すが，質問応答システムでは，文書単位の回答を行うのではなく，ユーザの質問に対する答えをずばり，数単語，1 文で提示するシステムである．本稿で試作した質問応答システムは，テキストからではなく，音声文書中から解答を探しだす．正しい解答を見つけるためには，質問内容の把握（誰，いつ，など），音声文書中の各単語に対する固有表現の付与が必要となる．質問内容分類は簡単なルールに基づいて行ない，固有表現の付与は既存の固有表現タガー，クラスベースの言語モデルによる直接推定により行なう．最後に，実際の音声質問を用いた応答実験では，固有表現タガーを用いるよりも，言語モデルを用いた手法の方が固有表現の抽出率は良くなり，質問応答の精度も向上した．

キーワード 質問応答，音声文書，音声質問

Speech-based Question Answering System for Spoken News Documents

Hiromitsu NISHIZAKI† and Seiichi NAKAGAWA††

† Interdisciplinary Graduate School of Medicine and Engineering,
University of Yamanashi
4-3-11, Takeda, Kofu, Yamanashi, 400-8511
†† Information and Computer Sciences, Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, 441-8580

E-mail: †nisizaki@ccn.yamanashi.ac.jp, ††nakagawa@slp.ics.tut.ac.jp

Abstract This paper describes a Japanese spoken question answering system which can find out an answer for spoken documents. In the question answering system, the question type extraction from an input and the named entity extraction from spoken documents are important to estimate an answer. So, we investigated two methods for extracting the named entity for spoken documents. One method is to use a Japanese NE tagger against a recognizer's output of spoken documents. The other is to extract directly NE by using a recognizer with a class-based language model. In an experimental result, we confirmed that using the class-based language model is more robust for OOV words and mis-recognized words than the NE tagger.

Key words question answering, spoken documents, spoken query

1. はじめに

現在の情報検索システムの多くは，ユーザが入力したクエリーに対する回答として文書全体を返す．しかし，実際にユーザが必要としているのは，5W1H や宣言知識などその文書に含まれている一部分であることが多い．たった，1 単語かもしれないし，あるいは 1 行に渡るかもしれないが，文書全体から自分が欲する答えを探し出すのには比較的労力がかかる．情報検索システムの次の段階として，質問応答システムが注目を浴びている．質問応答システムでは，文書単位の回答を行うのではなく，ずばりユーザの質問に対する答えを，数単語，1 文で提示するシステムである．海外では，アメリカ NIST の TREC のプロジェクトの一つに QA タスクがあるし，日本でも NTCIR が QA タスク(注1)を開催していることもあり，今後この分野の研究に拍車が掛かることが予測できる．本稿では，音声文書検索の発展として，音声文書データを対象とした音声入力型質問応答システムについて述べる．

一般的に，質問応答システム（テキスト入力，テキスト検索）では以下に述べるような構成要素が必要となる．

- 質問解析部
- 情報検索部
- 解答抽出部

質問解析部では，入力した質問を解析し，質問タイプ，検索語（検索用のキーワード）抽出などを行う．質問タイプとは,『場所』『人名』『組織名』など固有表現を使うことが多い．例えば,「アメリカの第 43 代大統領は誰ですか？」という質問に対しては，質問タイプは『人名』となる．最低限，質問タイプの推定と検索用キーワードの抽出を行えば，質問応答システムは動作する．しかし，高精度に解答を推定するために，文献 [1] では補助語や単位語，意味カテゴリの抽出も行っている．情報検索部では，質問解析部で抽出したキーワードを用いて解答が含まれている文書，段落，文を検索する．文書単位で検索を行うと，比較的大きな文章集合から，解答を推定しなければならないし，逆に文単位での検索では，解答が含まれていないことがあるので，段落単位での検索（パッセージ検索）を行うのが一般的である．パラグラフの検索は，情報検索で用いられている一般的な技術（TF･IDF 法など）を用いて行い，検索語との類似度が高い上位 10〜20 個のパラグラフを選択する．

解答抽出部では，検索部で得られたパラグラフ中の文章から解答位置を推定する．解答抽出の方法としては，質問タイプに合致する固有表現の単語（例えば，質問タイプが人名であれば，解答は人名を抽出）を選択することになるが，当然複数候補得られる可能性がある．解答候補の選択基準としては，検索用キーワードからの距離（単語数）を用いたり [1]，パラグラフを検索語とのマッチング数などを基準にしてさらに文を絞り込み，解答部分を推定する方法 [2] などが提案されている．

音声質問，音声文書データを対象とする場合，音声認識誤りや未知語の問題が立ちはだかる．今回は特に，音声文書中の単語に対する固有表現の抽出方法について，既存のタガーを用いる方法を比較検討した．クラスベースの言語モデルにより直接推定する方法を用い，実験の結果，固有表現のクラスと単語を混ぜ合わせたハイブリッド言語モデルを用いることで未知語に対してロバストな結果を得た．

(注1)：http://www.nlp.cs.ritsumei.ac.jp/qac/

2. 音声質問応答システム

質問応答システムにおいて，音声入力で質問を入力し，テレビニュース放送などの音声文書から自動的に書き起こしたデータベースから質問の答えを抽出する場合を考える．本研究で構築した音声質問応答システムの概念図を図 1 に示す．通常の音声文書検索の場合と同様に，音声認識システムで答えを見つけたい音声文書集合を認識し，書き起こしデータベースを作っておく．音声入力した質問は同じく認識システムで認識され，認識結果から質問タイプとキーワードを抽出する．質問タイプとは，入力した質問語がどういった答えを求めているかというもので，質問応答システムでは非常に重要なファクターである．質問タイプとしては，人名，地名，組織名，日付，時間，金額，人数などが考えられる．キーワードを使って音声文書集合から答えが含まれていそうな記事を検索し，検索された文書集合から質問タイプに合う部分（解答）を決定する．

音声文書，音声入力質問を対象とする場合は，どうしても音声認識での問題（音声認識誤り，未知語問題）というのが存在する．自然言語処理の観点からシステムを構築する方法は先行研究でいろいろと行われているので，今回は特に，音声入力，音声文書を扱う場合の音声処理の観点からシステムを構築することに重点をおき，システムの開発を行った．

2.1 音声質問応答システムの要素技術

音声文書中から解答位置を推定するための，音声質問応答システムに必要な要素技術について述べる．

音声質問からの質問タイプの決定：

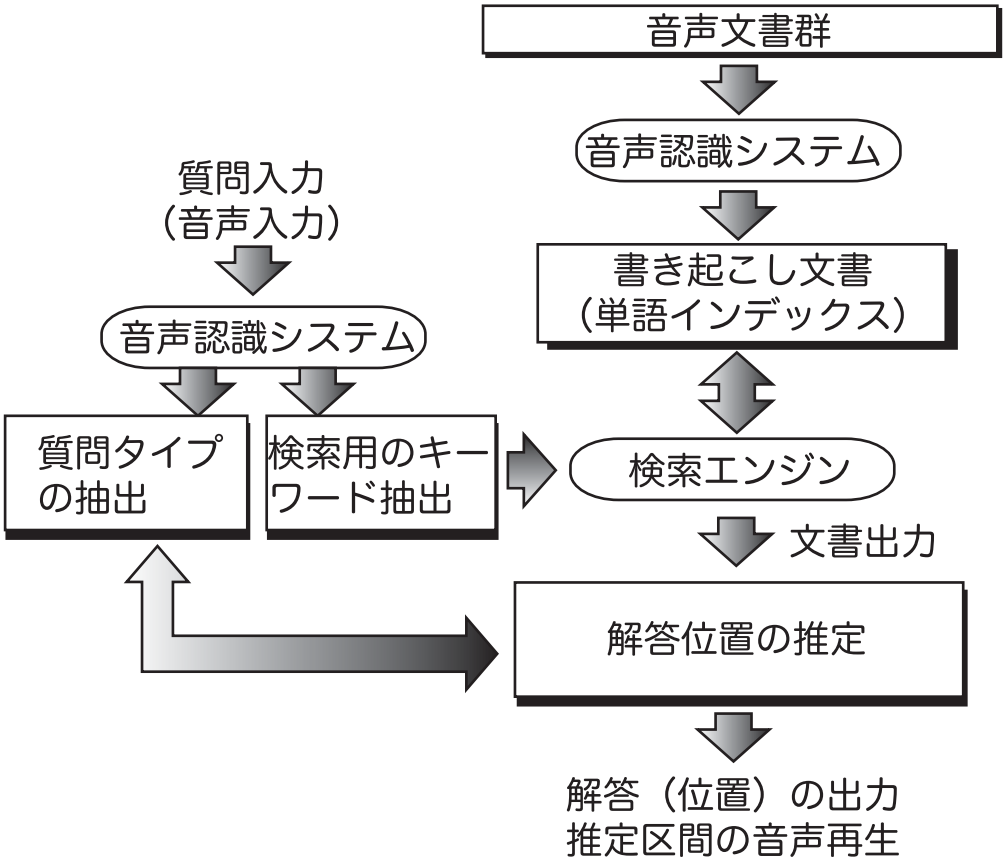

図 1　音声質問応答システムの概念図

音声入力した認識結果から質問タイプを決定しなければならない．一般的に行われているのは，質問のタイプ毎にルール（パターン）を設定しておき，そのルールとマッチングすることで質問タイプを決定する．たとえば，「〜はどこにありますか？」という質問で，『どこ』という単語(注2)が入っていれば，地名と組織名が対応する質問タイプということになる．1 段階目として認識結果に対してこの処理を施す．質問タイプは IREX(注3) で定義された固有表現 8 種類 (組織，人名（固有物），地名，日付，時間，金額，割合，人数) を用いる(注4)．

当然，音声認識誤りが含まれているため，

- 質問タイプを決定するキューとなる単語がうまく認識できていない
- 実際の質問とは違った質問タイプのキューとなる単語が湧き出てくる

という問題が出てくる．今回は単純にルールベースで質問タイプを検出する．

音声質問からキーワード群の決定：

音声文書を検索するのに必要なキーワード群を抽出する．ここで問題になるのは，

- キーワードの認識誤り（脱落，挿入，置換）
- 未知語

である．認識誤りの問題に関しては複数の認識システムの出力の混合を用いることで認識率の向上が図れる [3].

キーワードを用いた文書検索：

音声質問の認識結果からストップワードを除去し（名詞以外），文書検索用キーワードを抽出する．それを用いて文書集合から関連のある文書を検索する．文書は TF・IDF に基づく類似度順に出力する．今回は類似度が 1 位の文書のみを用いて，次に述べる解答部分推定を行う．

検索された文書から解答部分の推定：

検索エンジンで検索された音声文書から答えの含まれている部分を推定する．本システムでは，単純に検索キーワードからの距離（単語数）を基準に考える．キーワードからの平均距離が一番近くて，質問タイプと合致している部分を質問に対する解答候補として出力する．ここでの問題点として，

(注2)：こういった質問タイプの決定に用いる重要な単語をキュー単語と呼ぶことにする．
(注3)：情報検索と情報抽出のワークショップ．
http://www.csl.sony.co.jp/person/sekine/IREX/
(注4)：ただし，8 種類だけでは足りないので，適時追加していく必要はある．

- 音声認識誤りにより，解答が正しく認識されているとは限らない
- 音声認識によって得られた単語に対して，正しい固有表現が付与できるか
- 未知語の扱い

ということがあげられる．音声文書の場合は，テキスト文書の場合と異なり，解答候補のある位置を同定することで，その部分の音声を聞く（もしくは映像を見る）ことができるため，質問タイプさえあっていれば音声の誤認識の問題はない．たとえば，日付の間違い，時間の間違い，人名同士の間違い(注5)などは問題ない．しかし，質問タイプで解答位置を検出する場合は，質問タイプを決定する単語が全く違う単語に認識されてしまうと解答位置を検出できない．そこで，評価に関しては正解単語が含まれている位置を正しく推定できたかどうか，という判定を行う方法を採ることにする．

2.2 音声質問

このクエリーから質問タイプの検出を行う．質問タイプ決定は，単純なパターンマッチングにより決定する．質問タイプは，IREX の固有表現リストの 7 種類（＋ 1）で『人名<NAME>』『組織<ORG>』『地名（場所）<LOC>』『時間<TIME>』『日付<DATE>』『金額<MON>』『割合<RATE>』『人数<PERS>』である．表 1 に示すようなパターンを用いてマッチングを行う．

2.3 音声文書中の固有表現抽出

QA タスクを行うのに関してどうしても必要となる音声データベース中の固有表現の識別であるが，これを行うために，

- 音声文書の認識結果に対して，既存の固有表現タガー（三重大の『NExT』[4] (注6)）を用いる方法，
- 固有表現クラスと単語混合のハイブリッド言語モデル（LM）を用いて検出する方法，

を試みた．1 番目の方法は，通常の単語言語モデルを用いて認識した結果をタガーに入力して固有表現を付与する．この際，誤認識問題が発生するが，無視して処理を行う．

2 番目の方法の言語モデルに対しては，固有表現のみをクラス化したクラス・単語混合のハイブリッド言語モデルを作成し，どのくらい固有表現が判定できるかの認識実験を行った．単語→クラス，クラス→単語，クラス→クラス，単語→単語間の遷移確率は通常通り学習し，クラスから各単語への確率は，そのクラスに属する単語間で等確率になるようにした．クラスとしては，人名／地名／組織名のみ 3 クラスで行った．時間や日付，金額などの数値的要素が高い固有表現は，簡単な規則を用いてまとめあげるようにしている．また（連続）数字の後ろに” 時”，” 月”，“日”，“年”，“円”，“人”，“%” などが来るとそれに対応した固有表現を付与する．

3. 質問応答実験

3.1 データベースと質問要求

解答を見つけるためのデータベースとしては，音声文書検索の実験でもちいてきた NHK 音声データベース（1996 年 6 月 1 日〜7 月 14 日）を使用する．

入力する質問は被験者 2 名により作成した 50 個質問を使用する．それを男性話者 2 名が読み上げて収録した音声データを音声質問要求として使用した．なお，質問文は文献 [5](注7)や NTCIR の QA タスク(注8)で使用されている質問文と同等程度の複雑さである(注9)．

3.2 音声質問の認識と質問タイプ検出

収録した音声質問を，音声認識システム『Julius』[6] を用いて認識を行った結果（特徴量は 16KHz サンプリング，25ms ハミング窓，10ms シフトの MFCC，音節 HMM，言語モデルは NHK 汎用原稿），表 2 の認識率が得られた．

表 2　音声クエリーの単語認識率 [%]

Cor.	Acc.	Sub.	Ins.	Del.
86.6	84.4	10.6	2.3	2.8

次に，このクエリーから質問タイプの検出を行う．質問タイプは前述した 8 種類である．表 1 に示した簡単なルールを用いて認識結果に対する質問タイプを検出する．この方法で質問タイプを決定した結果，正しく検出できた割合が 64%，検出不可能だったものが 36%，誤って検出したのは 0% という結果になり，音声認識誤り，特に質問タイプを検出するキューとなる単語の誤認識により，全体で 64%の質問文しか正しく質問タイプを検出することができなかったが，誤った質問タイプを検出することはなかった．既存の質問応答システムではなんらかの基準により質問タイプの検出を行

(注5)：この場合だと人名の正しい漢字表記を得ることは無理．

(注6)：ルールベースのタガー．タグつけの性能としては，IREX ワークショップ参加チームの平均的な精度が得られている．http:www.ai.info.mie-u.ac.jp/ñext/next.html

(注7)：文献 [5] には質問 50 文がリストアップされている．

(注8)：ホームページに質問の一部が例として挙げられている．http://www.nlp.cs.ritsumei.ac.jp/qac/format02.html

(注9)：質問の例：「岡山県の邑久町で何人の子供が大腸菌による食中毒にかかりましたか？」,「国際捕鯨委員会の年次総会はイギリスのどこで開かれましたか？」

表 1　質問タイプ決定用のルール

パターン（正規表現）	対応するタイプ	質問パターン例文
.*だれ.*	人名	～したのは誰ですか？
.*名前.*	人名，組織	～の名前は？
.*いつ.*	時間，日付	～したのはいつですか？
.*どこの<ORG|LOC>	組織，場所	～したのはどこの国ですか？
		～したのはどこの会社ですか？
.*どこ.*	場所	～はどこにありますか？
.*何という<ORG|LOC>.*	組織，場所	～は何という地名でしたか？
		～は何という工場でしたか？
.*<ORG|LOC>は何と*	組織，場所	～している大学は何と言いますか？
.*何<TIME|DATE|MON|RATE>.*	時間，日付，金額，割合	何時，何日，何円，何%…

い，解答位置の推定を行っているが，本実験では，質問タイプの検出ができなかった場合（つまりどのルールも適応できなかった場合），検出不可となり答えの推定ができないことになる．質問タイプ検出の失敗の原因は，検出の手がかり（キュー）となる単語の誤認識である．キュー単語の認識率をあげる方法として，複数の認識システムでクエリーを認識することで，キューの認識率を向上させる，などの処理が必要である．また，2 段回目の処理として直接音声から想定されるキューをスポッティングする方法などが考えられる．

3.3　音声データベース中の固有表現抽出実験

NHK 音声データベースに対する固有表現の抽出実験を行った．音声データベースの認識には音声認識システム Julius（特徴量は 16KHz サンプリング，25ms ハミング窓，10ms シフトの MFCC，音節 HMM）を用いている．タグ付けの方法としては次の 2 通り：

- 認識結果に対して，固有表現タガー『NExT』を用いて固有表現を付与する．この場合，誤認識は無視することにする．
- 固有表現クラスと単語混合のハイブリッド言語モデルを用いて検出する．

の方法で行う．

50 個の質問文に対する解答のみに関して，上記の 2 通りの方法で固有表現の検出率を求めた．先にも述べたように，音声データから解答をマイニングする場合は解答の位置を推定できれば，そこの音声区間の音声を聴くことによって真の解答を得ることができる．その考え方から，認識された単語表記が正しくなくても，質問に対する正解単語（列）が含まれている区間に対して正しい固有表現が振られていれば良いという評価基準で固有表現の検出率を求めた．50 の質問文に対する解答位置はデータベース中に全部で 144 個所存在する．抽出結果を表 3 に示す．

テキスト文書に対して NExT でタグを付与した場合は，144 個の解答部分に関しては検出率は

表 3　固有表現の抽出実験結果 [%]

	テキスト文書	音声文書	
		NExT	ハイブリッド LM
検出率	100	57.6	70.1

100%である．音声認識結果に対して NExT のような既存のタガーで固有表現を付与しようとすると，音声認識誤りため正解のタグを付与することができないことがある．ハイブリッド言語モデルを用いることで，必要以上に固有表現のタグが付与されてしまう（検出の適合率が小さい）が，比較的再現率は高いと思われるので，誤認識が発生しても比較的同じ品詞の誤りとなることが多く，正しく固有表現を付与できる．また，解答が未知語である場合でも，認識された単語は解答ではないが，固有表現としては正解という場合もありうるので，使っているユーザが音声を聞けるという前提では，音声データにおいて解答候補の位置が正しく推定できていれば問題はない．また，ハイブリッド言語モデルを用いて固有表現を抽出する利点として，未知語に頑健で あることが実験結果からわかった．通常，認識辞書に含まれない単語，つまり未知語は，必ず他の登録単語に置き換わって認識されてしまう．この単語が未知語の固有表現と同じクラスであれば問題ないが，ほとんどの場合でまったく違うクラスになってしまう．今回の実験で，ハイブリッドな言語モデルを用いることで，単語は異なるが未知語の固有表現のクラスを正しく検出することが出来ている．

3.4　音声質問応答

これまで述べてきた，音声質問からの質問タイプ検出とデータベースの固有表現の抽出技術（ハイブリッド言語モデル）を用いて質問応答実験を行った．

処理の流れは，音声質問タイプの検出→検索語の抽出→文書検索，検索語との類似度がもっとも良かった文書を一つ選択→解答位置推定という流れである．

表 4　音声質問応答実験結果 [%]

	テキスト文書	音声文書
テキスト入力質問	75	50
音声入力質問	48	36

50 個の音声質問を用いた応答実験結果を表 4 に示す．音声の質問に対して，音声文書群から解答を推定する実験結果では，36%という結果になった．音声質問の場合は，50 個のうち 64%の質問のみ正しく質問タイプの判定が行われている．これらの質問に対しては 56%の精度で正解位置を特定できている．テキスト入力質問の場合は 100%質問タイプの特定ができているが，文書中から解答位置を推定する方法が非常に単純であるため，テキスト文書中を用いても精度が 2 割強低下する．音声文書中から推定する場合はさらに応答精度が低下し，約半分の精度しか得られていない．

今回の実験では，質問タイプが検出できなかった質問（36%）は解答が得られないという条件であったので，質問タイプをどれか一つに決定するということをやれば，質問応答の精度の向上は図れるはずである．また，ハイブリッド言語モデルでは，不必要な固有表現クラスが多量に湧き出してしまい，解答位置の精度が悪化している．今回はクラス言語モデルという単純な方法で固有表現の抽出を行ったが，文献 [7] などの先行研究で提案されている音声データからの固有表現の抽出法を用いることにより，さらに音声データからの固有表現抽出の精度上昇が見込める．さらに，今回は情報検索エンジンで得られた，検索語ともっとも類似度が良かった文書のみを解答位置の推定に用いた．当然，この文書に解答が含まれていなければ，正解の解答位置を見つけることはできない．質問に対する解答が含まれている文書は 1 つとは限らない（今回の実験では 1 つの質問に対して平均 3 文書で解答が含まれている）．正解の文書がうまく検索されてきたとしても，正解位置に間違った固有表現が付与されていることも十分考えられるので，検索エンジンで得られた上位 N 文書を用いることによって，さらに解答位置の推定精度を向上させることができると思われる．

4. おわりに

本稿では，音声入力型質問応答システムについて述べた．現在盛んに行われているテキストレベルの質問応答システムとは異なり，音声で質問を入力し，その解答を音声文書中から推定する方法を提案した．

質問要求も解答をマイニングするデータベースも両方音声であるため，音声特有の問題が起きる．質問応答システムでは，質問中からの質問タイプの特定と，データベース中の固有表現の抽出が精度に大きく影響する．

音声質問の音声認識結果から単純なルールを用いて質問タイプを検出する方法では，64%の検出精度しかえられなかった．この精度を改善する方法として，複数の認識システムで音声質問を認識することで，質問タイプ検出のキューとなる単語の認識率を向上させる，2 段回目の処理として直接音声から想定されるキューをスポッティングする方法などが考えられる．本実験の段階では質問タイプの検出が失敗すると処理が終わってしまうので，これらの方法を用いてタイプの決定を行う処理を追加する必要がある．

固有表現の抽出に関しては，認識結果に対して既存の固有表現タガーを用いる方法，固有表現をクラスとして組み込んだ言語モデルを利用する方法を用いた．実験の結果，テキストを対象とした場合によく用いられるタガーによる方法では，誤認識に弱く，言語モデルを使って直接推定した方が，誤認識や未知語に対して比較的ロバストであることがわかった．

質問応答システム全体の性能としては，50 個の質問に対する解答位置の推定精度は 36%であった．今後は，個々のモジュールの精度改善を行い，全体の精度の改善を図る予定である．

文　献

[1] 佐々木裕, 磯崎秀樹, 平博順, 平尾努, 賀沢秀人, 鈴木潤, 国領弘治, 前田英作. SAIQA：大量文書に基づく質問応答システム. 情報処理学会研究報告, 2001-NL-125-12, pp. 77–82. 情報処理学会, 2001.

[2] A. Ittycheriah, M. Franz, A. Ratnaparkhi W.J. Zhu, and R. J. Mammone. Question answering using maximum entropy components. In *Proc.of NAACL2001*, pp. 33–39, 2001.

[3] 小玉康広，渡邊友裕，宇津呂武仁，西崎博光，中川聖一．SVM を用いた複数の大語彙連続音声認識モデルの出力の混合．日本音響学会，春期講演発表会演論文集 I，3-4-11，pp.151–152，2003.

[4] 桝井文人, 鈴木伸哉, 福本淳一. テキスト処理のための固有表現抽出ツール next の開発. 第 8 回年次大会発表論文集, P2-9, pp. 176–179. 言語処理学会, 2002.

[5] 佐々木裕, 磯崎秀樹, 平博順, 廣田啓一, 賀沢秀人, 平尾努, 中島浩之, 加藤恒昭. 質問応答システムの比較と評価. 電子情報通信学会技術研究報告, NLC2000-24, pp. 17–24. 電子情報通信学会, 2000.

[6] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 嵯峨山茂樹, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏. 日本語ディクテーション基本ソフトウェア (99 年度版). 日本音響学会誌 (技術報告), Vol. 57, No. 3, pp. 210–214, 2001.

[7] F. Bechet, A. Gorin, J.Wright, and D.Hakkani Tur. Named entity extraction from spontaneous speech in how may i help you? In *Proc.of ICSLP2002*, pp. 597–600, 2002.