Vol.2010-SLP-84 No.21
2010/12/21

QA システムのための音声入力インターフェース

松田繁樹[†1] 林輝明[†1] 大竹清敬[†1]
Stijn De Saeger[†1] Istvan Varga[†1] Yulan Yan[†1]
風間淳一[†1] 磯谷亮輔[†1] 河井恒[†1]
鳥澤健太郎[†1] 中村哲[†1]

思考や行動の幅を広げるためのツールとしての QA システムとその音声入力インターフェースについて紹介する．いつでも・どこでも日常のふとした思いつきから思考や行動のオプションを広げるためにスマートフォンを用いた音声入力インターフェースを持つ QA システム「一休」を開発した．QA システムは WWW から抽出したさまざまな意味的関係パターンのデータに基づいて回答を検索する．本稿ではドメインに依存しない音声による質問を可能とする為，言語モデルを大規模 WWW テキストコーパスから作成し，50 万を越える語彙を持つ音声認識による音声入力インターフェースを構築した．

Speech interface for QA systems

Shigeki Matsuda,[†1] Teruaki Hayashi,[†1]
Kiyonori Ohtake,[†1] Stijn De Saeger,[†1]
Istvan Varga,[†1] Yulan Yan,[†1] Jun'ichi Kazama,[†1]
Ryosuke Isotani,[†1] Hisashi Kawai,[†1]
Kentaro Torisawa[†1] and Satoshi Nakamura[†1]

This paper introduces the "Ikkyu" question answering (QA) system and its speech interface. Ikkyu can be used both as a search engine proxy and a tool for innovation support. To be able to use Ikkyu anytime and anywhere, we employ smartphones as an interface to the QA system. The QA system uses semantic relation acquisition techniques to retrieve many possible answers for its input questions from the Web. Touching the words that were output as answers allows the user to check the original information source of various useful answers. In order to realize a fully open-domain speech interface, the system's language model was constructed from a large collection of Web documents and contains a lexicon of more than 500k words.

1. はじめに

本稿では，思考や行動の幅を広げるためのツールとしての QA システムとそのための音声インターフェースについて紹介する．我々の研究グループでは，これまで，単語間の意味的関係によって概念を派生させ，利用者の要求にしたがって成長させることができる概念辞書をはじめとして，単語の意味的な分類のための言語資源やツールを整備し，公開してきている．今回，単語の意味的関係を抽出する技術を応用し，オンデマンドで単語間の意味的関係をリアルタイムに抽出し質問に回答する音声質問応答システム「一休」を開発した．「一休」は，iPhone 等のスマートフォンを持つユーザが，自宅，街中や，観光名所などいつでも，どこでも利用できることを想定している．ふとした思い付きや疑問などをすばやくシステムに質問できるようにする為，音声入力インターフェースを導入した．

「一休」は，WWW を対象として質問の回答を与えるシステムであるため，その音声認識システムは様々な語彙を認識する大語彙音声認識が必要である．「一休」は，当機構が開発している ATRASR を音声認識デコーダとして用い，50 万を越える語彙を持つ言語モデルによって音声入力を行うことができる．また，騒音のある屋外での利用を想定しているため，導入する音声認識システムは，雑音に対して高い頑健性を持つ必要がある．パーティクルフィルタを基礎とした雑音抑圧，及び雑音重畳音声を用いて推定した音響モデルを併用することにより耐雑音性能の改善を行った．

2. QA システム概要

「一休」は，WWW を対象として質問の回答を与えるシステムである．従来の Factoid 型の QA システムとは異なり，WWW 上に記述された膨大な回答となる表現に触れることができるシステムである．古典的な QA システムは，「世界で一番大きい湖はどこですか」といった正解がひとつしかない質問に対して適切な応答を生成する．一方，本システムでは，「円高の原因は何ですか」のようなひとつの正解よりは，どのような事を原因としてあ

†1 情報通信研究機構
National Institute of Information and Communications Technology

Vol.2010-SLP-84 No.21
2010/12/21

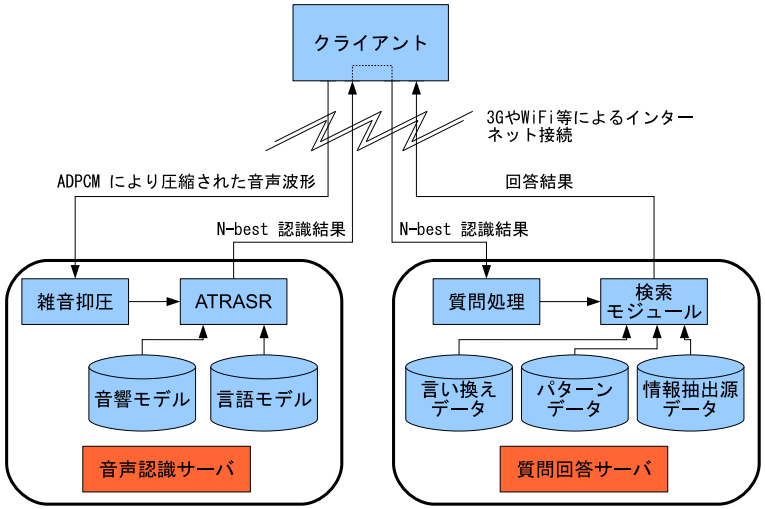

図 1 QA システム「一休」の概略図

げているかを網羅的に知りたい質問に対して,「貿易黒字」,「バブル崩壊」などといった様々な回答を与えることができる．さらに，これらの単語がどこのページから抽出されてきたかを表示させることができるので，オリジナルのページを確認することで，その情報の信頼性を確認したり，その他の関連情報に容易にアクセスできる．また，本 QA システムでは，2 単語間の関係は明確に区別されるので「円高の原因」と「円高の帰結（結果）」は完全に異なる結果を返す．つまり,「円高は何を引き起こしますか」という質問と「何が円高を引き起こしますか」が区別される．既存の検索エンジンを使用して前述のような質問の回答を得ようとしても両者が混在したページを確認することになるため，質問に対する回答をピンポイントに得ることは困難である．

この QA システムを用いることで，思考，行動のオプションを広げることができると考えられる．日常のふとした思いつきから，思考や行動のオプションを広げるためには，いつでも・どこでもシステムを利用できることが望ましいと考え，スマートフォン上にシステムを実装し，さらに音声インターフェース用意することで，入力の煩わしさを解消した．

2.1 質問回答サーバ

質問回答サーバで用いられる技術は，基本的に，我々がこれまでに開発してきた概念辞書を構築するための技術の応用である．概念辞書を構築するために，大規模な WWW コーパスを利用して，2 単語間の関係を抽出する手法を開発した[1)]．これを関係抽出器といい，関係抽出器によって抽出された 2 単語間の関係をパターンと呼ぶ．たとえば,「タミフルでインフルエンザが治る」という表現からは,「X で Y が治る」というパターンを抽出することができる．抽出されたパターンをその変数のインスタンスとともに保存しておき，これを用いて回答を与える．

起動ロゴ　使用方法説明画面　認識中画面　検索中画面（認識結果n-best使用時）　検索結果（回答）表示画面

図 2 iPhone クライアント動作時画面

抽出されたパターンをそのまま検索するだけでは，十分な回答を与えることができないので，パターンの変数のうち，同じものが数多く共有される別のパターンを言い換えパターンとみなし，回答を与えるパターンの対象を広げる．たとえば,「X で Y が治る」の変数のうち，X=ステロイド剤，Y=アトピーや，X=アスピリン，Y=頭痛といったインスタンスが別の「X で Y を治療する」というパターンでも多く出現するならば，このパターンは言い換えパターンである可能性が高い．このように言い換えとなるパターンも事前に獲得しておく．

質問が入力された時の質問回答サーバの動作の概略は次のとおりである．

(1) 質問を解析し，質問に回答を与えるパターンを推定する．たとえば,「アトピーの原因は何ですか」に対して「X が Y の原因（Y=アトピー）」というパターンが回答を与えるパターンであると推定する．

(2) パターンのデータベースに対して質問に回答を与えるパターンを検索し，回答の候補を求める．この場合，言い換えパターンも適用し，網羅性を高める．たとえば,「X が Y の原因（Y=アトピー）」ばかりではなく,「X が Y を引き起こす（Y=アトピー）」,「X が Y を誘発する（Y= アトピー）」や,「Y が X に起因する（Y=アトピー）」な

どといったパターンを検索し，回答候補となる X を求める．

(3) 回答候補を，頻度や，用いたパターンなどから評価し，ランキングを行い出力する．

2.2 システム構成

「一休」の構成を図 1 に示す．本システムは，音声を入力するクライアントと，音声認識を行う音声認識サーバ，質問の回答を検索する質問回答サーバとによって構成される．質問回答サーバが動作するためには，10TB 程度のディスクと，1〜2GB 程度のメモリが必要である．また，音声認識サーバを動作させるためには，現在使用している音声認識デコーダでは 4GB から 6GB 程度のメモリを必要とする．音声認識のクライアントは Apple の iPhone 用アプリとして実装されている．質問の結果は，音声入力に使用した iPhone，iPad 専用クライアント，もしくは PC の WWW ブラウザ上のいずれかに表示させることができる．

iPhone クライアントは，耳に近づけるとバイブレーションで録音の開始をユーザに知らせる．ユーザが話しおわり iPhone を耳から離すと，バイブレーションで録音の終了をユーザに知らせる．iPhone から 3G 回線もしくは，WiFi 接続などによるインターネット回線で音声が音声認識サーバに転送され認識される．認識結果は N-best のテキストとして iPhone クラアントに返される．iPhone クライアントは認識された N-best のテキストを質問回答エンジンに送信される．質問回答サーバは，回答結果を接続しているクライアントに応じて異なる形式で返す．iPhone の場合は，単語のリストを，そして，PC 上の WWW ブラウザの場合は，flash によるグラフィカルインタフェースを，iPad の場合は PC 上の WWW ブラウザと同等内容を html5 によって記述したものを返す．

iPhone クライアントの動作時画面を図 2 に示す．また，WWW ブラウザでの表示画面を図 3 に示す．

3. 音声認識システム概要

3.1 音声認識デコーダ

本 QA システムに音声入力インターフェースを導入するにあたって，当機構で開発している ATRASR を用いた．ATRASR は，単語バイグラムを用いて単語仮説と単語グラフを生成する第 1 パスと，単語グラフに含まれる単語仮説をプルーニングする第 2 パス，単語トライグラムによるリスコアリングの第 3 パスを行うことにより音声認識を実行する．また，本報告の言語モデルでは用いていないが，第 1 パスでは，単語バイグラムだけでなく単語長が可変の複合語やクラス化されたバイグラムを用いて認識することが可能である．本システムで対象とする約 50 万語彙の音声認識では，4GB から 6GB 程度のメモリを必要とする．

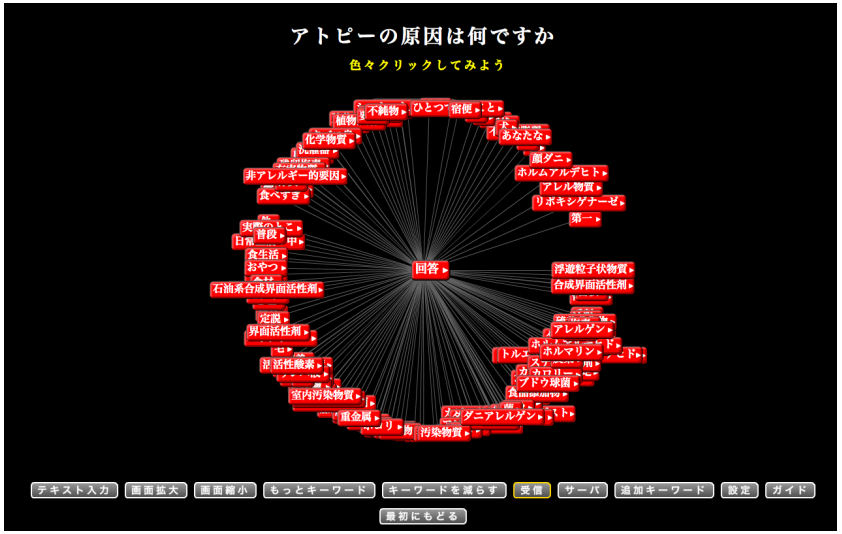

図 3 WWW ブラウザでの質問回答表示

3.2 WWW 言語モデルと辞書

本 QA システムは WWW に基づいたシステムであり，WWW 上の出現する単語を幅広く網羅するために，言語モデルも WWW から抽出したテキストに基づいて作成した．具体的には，Tsubaki コーパス[2] の 6 億ページから抽出されたテキストに対してフィルタリングを実行し，WWW コーパスを用意した．それに対して chasen を適用し，形態素解析を行った．chasen で使用した辞書は，ipadic-2.6.3 に基づきつつ，話し言葉に対してより頑健となるよう接続表の拡充，形態素辞書の追加を行ったものである．追加した形態素辞書は，Wikipedia の括弧内の読みの情報を利用して作成したものなどである．現在，形態素辞書の語彙サイズ（活用するものをすべて展開した大きさ）は 120 万ほどである．形態素解析に加えて，数詞と助数詞の読みをより正しくするために chawan を適用し，その結果を用いて言語モデルを作成した．

本稿で行った実験では 2 種類の言語モデルを用いた．ひとつは，もとの WWW コーパスからランダムにサンプルしてきて作成した言語モデルであり，もう一方は，文末の表現から疑問文と考えられる文に限定して作成した言語モデルである．前者を WWW ランダムモデ

Vol.2010-SLP-84 No.21
2010/12/21

表 1 言語モデルとテストセット

言語モデル	WWW 疑問文	WWW ランダム	テストセット
形態素数	163,108,013	165,751,181	22735
発話数	13,108,012	15,751,177	2500
平均 形態素数/発話	12.44	10.52	9.09
1-gram 数	455,381	654,483	1,550
2-gram 数	13,120,117	18,885,810	3,496
3-gram 数	45,855,339	60,755,977	4,127
テストセット PP (2-gram)	95.57	137.49	—
テストセット PP (3-gram)	59.69	85.85	—
OOV 率 (%)	0.73	0.68	—

ルと呼び，後者を WWW 疑問文モデルと呼ぶことにする．また，今回の実験では，共通の辞書などを用いずに，コーパスからそれぞれ直接生成した辞書を使用することとした．

3.2.1 WWW コーパスフィルタリング

WWW データを大語彙連続音声認識の学習データとして用いる試みは，たとえば，文献 3) や 4) など，WWW データが膨大になりつつあるころから行われてきている．我々も，それらの文献でとられた方法と類似したフィルタリングを適用し，言語モデルを作成している．Tsubaki コーパスから言語モデル作成までの流れを図 4 に示す．具体的には，次に示すフィルタリングを行う（1）共通フィルタリングでは，アルファベットのみからなる文や，記号類の削除，一部の表記ゆれ（ポケットテッシュ→ポケットティッシュ）やかな漢字変換のミス（彷佛→彷彿）などの置換を行う．(2）疑問文コーパスのため疑問文抽出は，文末表現などからルールベースで疑問文と判定できるものを抽出する．(3）解析・フィルタリングでは，chasen と chawan で解析された結果のうち，今回は，QA システムの入力部分を構成するため（a）指示詞と一部の連体詞（ここ，そこ，あそこ，この，など）と代名詞（私，あなた，など）を含む文，(b）数詞を含む文が発話されることが少ないだろうと考え，これらの形態素を含む文は学習コーパスから除くこととした．また，使用する形態素解析器の辞書サイズが大規模であることから，形態素解析結果のうち，未知語となるものを含む文は，辞書や言語モデルに含めるには不適切な表現が多かったので，これらの文も含めていない．ただし，表層文字列がカタカナによってのみ構成されてる未知語は含めることとした．その他，アルファベットによって構成される 32 文字以上の非常に長い形態素を含む文も同様に不適切な表現が多く見られたので，これらの文も排除することとした．

複数回出現する文に含まれる要素が全体として性能の向上に貢献せず，反対に性能を悪化させるものが多く含まれている可能性が高いことが予備実験の結果からわかった．そこで，

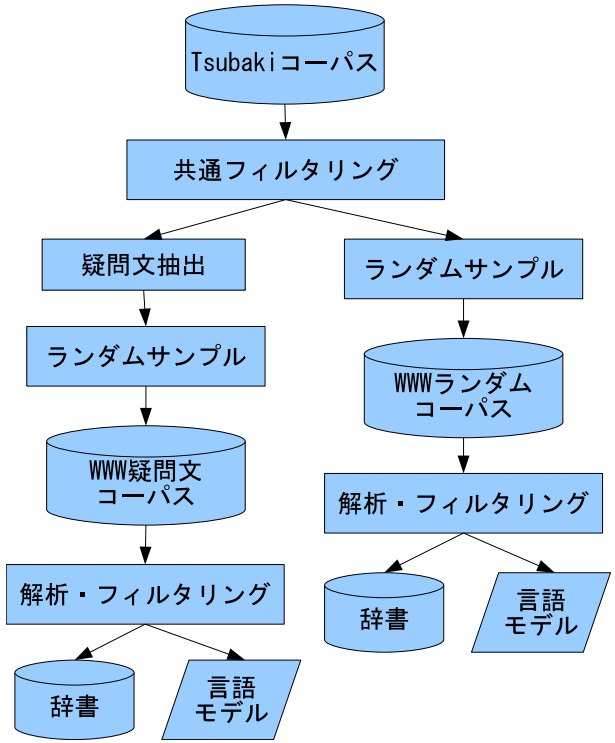

図 4 言語モデル作成フロー

いずれの言語モデルの学習コーパスも，複数回出現する文は，それぞれ一度だけ出現したものとして学習コーパスに含めることとした．

2 つの言語モデルならびにテストセットコーパスについて表 1 にまとめる．なお，両言語モデルを作成するために使用した形態素数は 1 億 5 千万形態素であるが，文頭と文末の特殊記号を計上している都合で，形態素数が異なっている．

3.3 音響モデル

本 QA システムは，屋外等で利用することを想定している．耐雑音性能を改善するため，入力された音声に対してパーティクルフィルタを用いた雑音抑圧[5] を行った．パーティクルフィルタによりフレーム毎に雑音スペクトルが推定され，ウィーナーフィルタを用いて入力

Vol.2010-SLP-84 No.21
2010/12/21

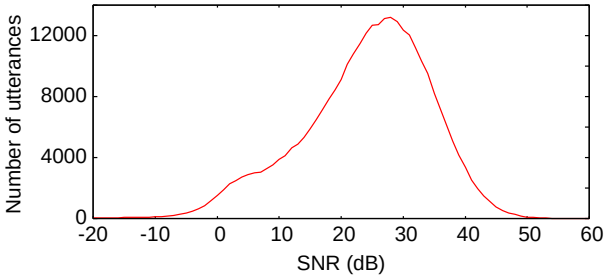

図 5 音声翻訳アプリ VoiceTra で収集された音声の SNR の分布

音声から雑音成分が除去される．さらに，表 2 に示す 17 種類の雑音を重畳した音声コーパスを用いて音響モデルを推定することにより，耐雑音性の改善を行った．

音声波形は，サンプリング周波数 16kHz，分析窓長 20ms，分析周期 10ms で分析を行い，MFCC 特徴量を抽出した．MFCC の音響特徴パラメータは，12 次元 MFCC，Δc_0，12 次元 ΔMFCC の計 25 次元である．使用した音素は，/ɴ, a, b, ʧ, d, e, f, g, h, i, ʒ, k, m, n, o, p, ʕ, r, s, ʃ, t ʦ, u, w, j, z/の 26 種類である．音響モデルの状態共有構造は，MDL-SSS[6] により生成した 5670 状態の HMnet を使用した．各状態の混合数は 10 である．性別に依存した 2 つの音響モデルを，ATR 旅行対話データベース TRA 及び，スケジューリング会話データベース APP，音素バランスの読み上げ音声 TRA-BLA と APP-BLA，更に当研究所で収録した高齢者音声を用いて推定した．総発話文章数は約 20 万文である．さらに，当機構より実験公開中の iPhone 用音声翻訳アプリケーション VoiceTra[7] で収集された約 70 万文を用いて教師無し適応を行った．一般化単語事後確率 (GWPP)[8] を用いて単語毎の信頼度を計算し，しきい値以上の信頼度を持つ音声区間を用い，音響モデル内の平均ベクトルのみに対して MAP 適応[9] を行った．

4. 音声認識評価実験

4.1 実験条件

音声認識用のテストセットを次の手順で作成した．

(1) 想定質問文の作成（約 800 文）．作業者 5 名に質問回答システムの概要を説明した上で，どのような質問を聞いてみたいかを一人あたり 100 文以上，記述してもらった．
(2) 男性 25 名，女性 25 名による読み上げ収録．一人あたり 100 文を想定質問文よりランダムに選択し，読み上げた．収録は，質問回答システムの音声認識インタフェースが iPhone 上に実装されていることから，iPhone 3GS を用いて行った．
(3) 想定質問文を言語モデル構築時と同様の手続きにて解析し，テストセットとした．

表 2 評価実験に使用した雑音

音響モデル推定に使用した雑音	
駅ホーム	駅前待合わせ広場
新幹線車内	空港ロビー
飛行機内飛行音	自動車内走行音
公共バス	田んぼ
競り市場	ファミリーレストラン
デパートエレベータホール	スポーツジム
スーパーマーケット	ホテルロビー
道路切断工事	オフィス受付
居酒屋	

評価用音声に重畳した雑音		
ファーストフード	商店街	駅コンコース

今回は，このうち，男性 12 名，女性 13 名，合計 25 名による 2,500 発話を用いて評価を行った．

実験公開中の VoiceTra で収集された音声の信号対雑音比 (SNR) の分布を図 5 に示す．SNR の計算には，音声認識結果に基き決定された音声区間の平均パワーと雑音区間の平均パワーの比を用いた．図 5 に示すように，約 25dB を中心に分布している．本評価実験では，SNR の分布の 90%をカバーする 10dB 以上の SNR(40dB, 25dB, 15dB, 10dB) の雑音を重畳して評価実験を行った．重畳した雑音の種類は，屋外での利用を想定し，表 2 の下段に示す 3 種類を用いた．雑音のチャネル特性は，iPhone 3GS 端末に調整した．辞書サイズは，表 1 の 1-gram 数に示す数と同じである．WWW 上の疑問文から作成した言語モデルは約 45 万語，WWW 上のテキストをランダム抽出した文章から作成した言語モデルは約 65 万語である．

4.2 実験結果

表 3 に実験結果を示す．表中の WER は単語誤り率，SER は文誤り率を表す．表中の 1-best の結果において，WWW 疑問文モデルは，WWW ランダムモデルよりも低い単語誤り率が得られた．タスクに適したテキストコーパスをフィルタリングにより抽出すること

表 3 音声認識実験結果

言語モデル	WWW 疑問文			WWW ランダム		
	1-best		20-best	1-best		20-best
	WER (%)	SER (%)	SER (%)	WER (%)	SER (%)	SER (%)
clean	13.38	54.00	32.64	15.60	57.08	33.72
25dB	12.63	49.60	26.64	13.62	51.58	27.64
15dB	14.17	53.83	31.32	15.87	57.30	33.28
10dB	19.82	62.54	44.36	21.57	64.59	45.20

で，低い単語誤り率が得られることを確認した．iPhone で収録した音声に 25dB の条件で雑音を重畳した評価音声に対する認識率は，クリーン音声の場合と比較して若干低い単語誤り率が得られた．15dB では若干の単語誤り率の増加，10dB では 7%程度の増加が見られた．本実験で用いた音響モデルは VoiceTra で収集された音声データを用いて適応されており，雑音レベルのミスマッチ等により認識誤りが増加したと考えられる．「一休」は，音声認識サーバから得られた 20-best の中からユーザが任意の認識結果を選択して質問回答サーバに送信することができる．そこで，20-best に対する文誤り率の評価を行った．WWW 疑問モデルを用いた場合，25dB で 26.64% の文誤り率であり，10 回の発話中 7，8 回程度の正解率が得られた．

本音声認識実験では，単語や文単位での認識性能の評価を行った．しかしながら，質問回答サーバは，文末表現や，一部の助詞の抜け落ち等の認識誤りに対して回答の品質を極端に劣化させるものではない．実際の使用感としてはスムーズな音声入力が可能となっている．今後は，ユーザ満足度等によりシステム全体の評価を行う予定である．

5. むすび

本稿では，QA システム及び，音声入力インターフェースについて述べた．ドメインに依存しない音声認識システムを構築するため，WWW 上のテキストコーパスから疑問文をフィルタリングにより抽出した文章から推定することで，約 50 万語彙を持つ言語モデルを用いた．また，スマートフォンを屋外で利用することを想定し，雑音抑圧処理及び，雑音重畳音声を用いて学習した音響モデルを用いた．音声認識実験から，25dB の評価音声に対して約 13%の単語誤り率が得られた．さらに，ユーザが任意の認識結果を選択して質問回答サーバに送信した場合を想定した 20-best での評価では，約 26%の文誤り率であった．

今後は，アンケート等を行うことにより，本 QA システム全体の評価を行う予定である．また，現在の質問回答サーバは，What, Who, Where, When を尋ねる短い質問にしか回答できないが，今後は，より長い質問や，How 型, Why 型の質問へも対応する．また，現状では，2 語間の関係を同一文内からしか抽出できていないが，日本語の場合，係助詞「は」などによって導かれる主題は，後続する数文に影響を与え，関係を持つ場合が多い．このような複数の文に存在する関係などにも対応し，網羅的に回答を発見できるようにする予定である．

参考文献

1) De Saeger, S., Torisawa, K., Kazama, J., Kuroda, K. and Murata, M.: Large Scale Relation Acquisition using Class Dependent Patterns, *Proceedings of the IEEE International Conference on Data Mining (ICDM'09)*, pp.764–769 (2009).
2) Shinzato, K., Shibata, T., Kawahara, D., Hashimoto, C. and Kurohashi, S.: Tsubaki: An open search engine infrastructure for developing new information access, *Proceedings of IJCNLP'08*, pp.189–196 (2008).
3) 伊藤克亘，秋葉友良：WWW は大語彙連続音声認識の学習データとして使えるか？，日本音響学会秋季研究発表会講演論文集，pp.131–132 (2002).
4) 西村竜一，長友健太郎，小松久美子，黒田由香，李　晃伸，猿渡　洋，鹿野清宏：Web からの音声認識用言語モデル自動生成ツールの開発，情報処理学会研究報告音声言語情報処理 35-8，pp.43–48 (2001).
5) Fujimoto, M. and Nakamura, S.: A Non-stationary Noise Suppression Method Based on Particle Filtering and Polyak Averaging, *Trans. IEICE*, Vol.E89-D, No.3, pp.922–930 (2006).
6) Jitsuhiro, T., Matsui, T. and Nakamura, S.: Automatic Generation of Non-uniform HMM Topologies Based on the MDL Criterion, *Trans. IEICE*, Vol.E87-D, No.8, pp.2121–2129 (2004).
7) VoiceTra アプリサポートページ：http://mastar.jp/translation/voicetra.html.
8) Soong, F.K., Lo, W.K. and Nakamura, S.: Generalized word posterior probability (GWPP) for measuring reliability of recognized words, *Proceedings of SWIM2004* (2004).
9) 中川聖一，越川　忠：最大事後確率推定法を用いた連続出力型 HMM の適応化，日本音響学会誌， Vol.49, No.10, pp.721–728 (1993).