中国語文型の特徴を利用した観光質問応答システム

胡　海青*　　任　福継*　　黒岩　眞吾*　　張樹武**

本論文では,回答精度及び頑健性の向上を目指し,ドメインを観光情報という中規模の領域に特定すると共に,統計ベースと浅い解析ベースを統合するQ&Aシステムの構成をする。インタフェースにおいて従来のテキスト入出力の上に，中国語の文型の特徴を基づき，言語モデル FSN(Finite State Network)、音響モデル HMM(Hidden Markov Model）と発音辞書を用い，特定領域向きの音声入出力も検討した。本稿では，形態素解析による中国語の言語特徴情報抽出とVSMモデルに基づき,よく聞かれる質問とその回答をデータベース化した質問応答データベースの回答検索と，旅行情報に関する文書の回答検索を統合するQ&Aシステムを構築した。特に文書検索においては，文書検索と文検索を統合することで，数文で的確な回答を行うことを目指す。また，評価実験を通じて本提案手法の有効性を検証した。

キーワード：質問応答，観光，文型特徴，音声認識，中国語

1. まえがき

質問応答(QA:Question-Answering)とは,自然言語で書かれた質問に対して,生の文書集合から適した答えを探し出す技術である。QA 技術に関しては，多くの研究が行なわれており，情報抽出，情報検索，自動要約，対話インタフェースなどの自然言語処理の各研究分野とも関連する技術である。質問における回答の絞込み手法としては,統計的な検索手法や浅い言語解析に基づくシステムが近年主流となっている。一方，対象分野を限定することによって，ドメイン知識の利用が容易となり，より高度な言語処理を行うことで，実用的な QA システムの構築が可能になるものと期待される。本研究では，日本の「観光立国」という政策の市場ニーズに応え，統計的な手法と解析ベースを統合した手法で日本の観光情報をドメインをとする中国語質問応答システムを構築する。即ち，よく聞かれる質問とその回答をデータベース化した知識ベース（以下は「質問応答データベース」という）の回答検索と，観光情報に関する文書検索を統合するものである。また，インタフェースにおいて従来のテキスト入出力の上に，中国語の文型の特徴を基づき，言語モデル FSN(Finite State Network)、音響モデル HMM(Hidden Markov Model）と発音辞書を用い，特定領域向きの音声入出力も検討した。

本論文の構成は次のとおりである。2.では，中国語質問応答システムの関連研究を紹介する。3.では，提案手法を述べるとともに QA システムの基本的な構成について述べる。4.では，音声入出力インタフェースのための音声認識と音声合成プロセスについて述べる。 5.では，構成したシステムの質問解析と回答候補の絞り込みプロセスについて述べる。6.では，提案手法に基づくシステムの評価実験結果を示すとともに,その結果について考察する。7.では，本論文のまとめと今後の課題について述べる。

2. 中国語 QA システムの関連研究

QA 研究への関心は,1999 年から TREC（Text Retrieval Conference）において QA-Track が新たに設定されたことをきっかけに高まりつつある。欧米や日本と比べて中国語の QA システムの研究は大変遅れていたが,近年では特に自然言語による中国語の QA システムについての研究が注目されるようになってきた。現在までにいくつかの中国語 Q&A システムが構築されているが，以上のような理由から主に統計検索手法や浅い言語解析等の手法を利用している。例えば,オープンドメインの代表的なシステムとして，中国科学院計算技術研究所の開発した大規模データベースに基づく「NKI 質問応答システム」（http://www.nki.net.cn）がある。また，哈爾濱工業大学応用ソフト研究室により開発された「QACAS」は，人手により作成された 73 の回答タイプルールを用いて構成された Web 文書を対象とした Q&A システムである。その他に，マサチューセッツ大学コンピューター科学部の Xiaoyan Li and W.Bruce Croft[1]により作られた「Marsha 中国語 Q&A システム」があり，英語文書に基づく QA システムの技術を中国語 QA システムに応用したものとなっている。一方,ドメイン限定システムとしては，旅行者を対象にする自然言語で登録された質問応答データベースの質問応答検索システムとして，中国科学院計算技術研究所の開発した「XIAOLINGTONG 質問応答システム」（http://159.226.40.18/ask/pub/）がある。また，もっと狭い領域を対象として，北京理工大学により開発された

*徳島大学大学院知能情報工学科, Graduate School of Engineering, The University of Tokushima
**中国科学院自動化研究所, The Institute of Automation Chinese Academy of Sciences

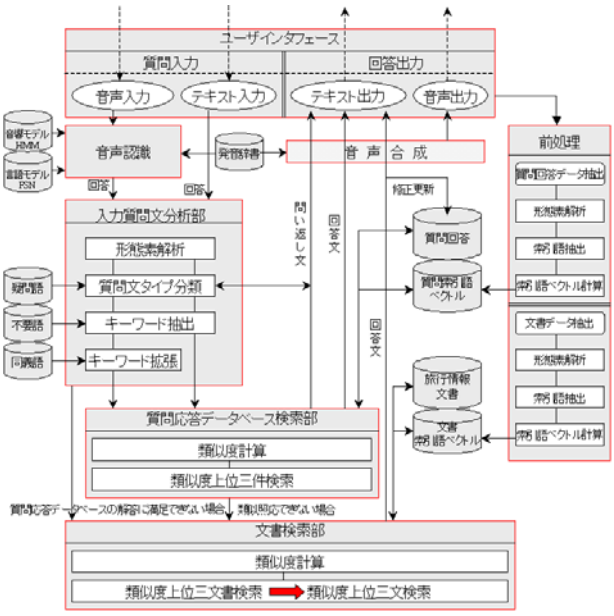

図１　システム構成図

金融ドメインに向けの商用システム「FAQAS」(銀行業務領域向け QA)及び中国科学院計算技術研究所の Shuxi Wang 等によって「紅楼夢人物関係 QAS」なども提案された。また, 2000 年，清華大学知能技術とシステム国家重点実験室により開発された清華大学キャンパスガイド「EasyNav」QA システムが実際に利用されている。以上紹介した中国語 QA システムの技術は，統計的な検索手法もしくは浅い言語解析に基づくシステムが主流である。金融ドメインに向けの QA システム「FAQAS」と紅楼夢人物関係(親族関係) の QA には，構文解析情報や浅い構造モード推理などの新しい技術を導入しているが，狭い固定的な領域に限られたものであり，観光情報のように比較的広く且つ更新が頻繁な領域に適用することは困難である。

3.　システムの基本構成

現在の中国語 QA システムは，NKI のように人手で作成した大規模なデータベースを必要とするためコストが高く，実際的なシステムの構築はかなり困難になる。これに対して，本研究では，基本的な出発点として，(1)データベースの収集しやすさを確保する,(2)人手のソースデータの取り扱いをできるだけ最小限にする,(3)新しい情報に応じデータベースを自動で動的に拡張できる，という三つの目標を立てた。本研究では，ドメインを観光情報という中規模の領域に特定すると共に，統計ベースと浅い解析ベースを統合する QA システムの構成を目指す。具体的には，形態素解析による中国語の言語特徴情報抽出と質問応答データベースによる応答, 及び VSM を用いた文書データベースからの応答生成により構成する。

一方で，QA システムが最終目標とする一言で明確な回答を示すことは技術的に困難であり，その精度は今だ低く使いものにならないといわざるを得ない。そこで，本システムでは，ユーザの質問文に類似した文書から類似度の高い回答文（回答を含めた文）数文を求める方式を採用する。

本システムは，日本への旅行者を対象として，中国語の自然言語による旅行情報の問い合わせに自動応答するものである。本システムの基本的な構成を図 1 に示す。本システムはユーザインターフェース部，音声認識と音声合成プロセス，質問文解析部，質問応答データベース部，文書検索部，前処理部といくつかのデータベースからなる。

4.　音声認識と音声合成プロセス

現在の質問応答システムの多くは，ユーザがテキストで入力したクエリーに対する回答を求める。近年の音声認識（ASR：Automatic Speech Recognition）技術は進展してきており，従来の情報検索タスクも音声入力に対応するように拡張されてきた。我々は,本システムでユーザからのテキスト入出力の上に，特定ドメインに向きの音声入出力の実用性も検討する。通常，多くの音声認識部は発音辞書、言語モデル（単語N-gram)、音響モデル(隠れマルコフモデルHMM： Hidden Markov Model)等のモジュールから構築される。

ところが，本システムが特定ドメインに限るということで,特定ドメイン向きの質問応答システムへの入力となる検索者の発話は、「定型的な表現となる質問文がよく使われる」という特徴が持っていることがわかる。そのため音声認識部では、定型表現が多いことに着目し，これらの高頻出定型表現には、記述文法としての有限状態ネットワーク（FSN：Finite State Network)は特定ドメインなどを対象としたタスクの限定された対話システムなどでは有効な方法である。したがって，本研究の観光システムで，音声認識部には，中国語の文型の特徴を基づき，発音辞書とHMMに基づくtriphone音響モデル(HMM) を用い，言語モデルとして単語N-gram言語モデルの代わりに，FSNで事前に定義された特定ドメインの文法を分析することによって、最も良い認識の性能を保証できるという構築が目指す。

本システムでは，観光情報に関する中国語の文型を分析し，高頻出定型表現を人手で抽出し観光ドメインへの有限状態文法を構成した。我々は中国語の文型の特徴を基づき，文法定義言語[2]を利用し，事前に 1 セットの観光分野関連の特定ドメインの文法を次のように定義した（一部の例)。

変数設計例:

```
$Polite = 请问 |(我 想 知道)| (告诉 我);
$Place = 景点 | 地方| 旅游景点|风景| ( 特色 景点)| (好玩 的);
$Where = 哪里 | 哪儿|(什么 位置)| (什么 地方)| (什么 方位);
$When = 何时 | (什么 时候)| (什么 时间)| (什么 季节)| 多久;
$Auxiliary = 可以 | 能;
$Modifier = 高 | 便宜| 贵;
$Which = 哪些 | 哪个;
$Traffic = 飞机 | 电车| 汽车| 巴士| 公交车| 地铁| 打的;
$TravelSite =日本 | 东京| 首都| 名古屋| 北海道| 富士山| 箱根;
$TravelSiteEvent = 温泉 | 温泉之乡| 有马温泉| 拉面| 樱花 ;
$TravelSiteMerit = 看 | 观看| 体验| 钓鱼| 吃| 洗|购物| 欣赏;
$TravelSiteCommon = 海拔 | 高度| 人口| 人| 电器;
文法設計例:
$GRAM1=(
 ( [$Polite] [去] [到] [从] $TravelSite [旅游][的 $TravelSite]
   ( 在 $Where ) |
   ( [ 到 $TravelSite ][坐|乘 $Traffic] 如何 (前往|去) | ( 到
     $TravelSite ) [呀|啊] )|
   ( 离 $TravelSite [很] 近 (么|吗) ) |
   ( [那儿|那里] 有 ( 什么|啥|$Which) [$TravelSiteMerit|有名|
     著名] [的] [$Place] [$TravelSiteEvent] ) |
   ( 好玩|好 [吗] 不好玩 ) |
   ( 的特色|$TravelSiteEvent (是 什么)|( 有 $Which) ) |
   ( [有名|著名] 的 $Place|$TravelSiteEvent (是 什么) | (有
     $Which)|(在 $Where ) ) |
   ( $When [去] [玩] 有意思 | 合适 | 最好) |
   ( [$TravelSiteCommon] 有 多|多少 [$Modifier] ) |
   ( [比 $TravelSite] 怎么样) |
   ( 有没有|有 [什么] [$Auxiliary] [$TravelSiteMerit
     $TravelSiteEvent] [的] [有名|著名|好] $Place | 注意事项) |
   ( 哪些 $Place 值得 去 [玩] ) |
   ( ($Where) 可以 $TravelSiteMerit $TravelSiteEvent )
 ) );
```

ただし，小括弧は一体性項目（括弧内の内容は一つの部分になる）を示し，縦線は代替性項目（いくつの中で一つを選べる）を示し、角括弧は任意性項目（オプション）を示す。

また，我々は特定ドメインの5kの単語の下に，この音声認識エンジンは1倍のリアルタイムで単語認識率が94%以上達することから，質問応答データベースの検索のため，よく現れる定型表現で聞かれた質問に対応することが可能であることがわかる。一方で,システムの頑健さを高めるため音声誤認識の場合，ユーザからテキストで修正及び再入力で回答の検索を行うことができる。

回答文音声合成について，本システムでは文章から検索された回答文は web ファイルから収集しているため，多くの未知記号を除去する必要がある。そこで，本システムでは回答文を形態素解析した文で音声合成をすることで，音声合成の質を上げている。

5. 質問解析と回答候補の抽出プロセス

〈5・1〉 質問文に対する解析処理

形態素解析：本システムでは，中国科学院計算技術研究所の開発した形態素解析システム ICTCLAS(Institute of Computing Technology Chinese Lexical Analysis System) [3]を用い形態素解析を行い，中国語の言語特徴により 39 種類の品詞タグ（例えば，一般名詞(n)，名動詞(vn)，名形詞(an)，普通動詞(v)，副動詞(vd)，助詞(u)，区別詞(b)，后接成分(k)等[4]）を付与した。

質問タイプの分析：HowNet Knowledge Database[5] から 55 の疑問詞を抽出した。これらの疑問詞により質問タイプを以下の 11 種類に細分化した。即ち，CAUSE（原因），TIME&DATE（時間&日付），LOCATION（場所&位置），PERSON（人名），METHOD（方法＆手段＆方式），QUANTITY＆AMOUNT（数量＆金額），DEGREE（程度），SUBSTANCE＆DEFINITION（実体＆定義），EVENT（事柄），REQUEST TONE（要請口調），OTHERS（その他）。

キーワードの抽出：本システムでは，形態素解析により、助詞，形容詞，及び一般的な副詞と動詞などの意味内容を直接表現しない単語(機能語)等を除いてから，また不要語データベースを用い取り除いたものをキーワードにする。本処理は後述する前処理過程と同じである。

キーワードの拡張：自然言語のゆらぎ（同じ意図が異なる語句で表現される）に対応するためキーワードの拡張が必要である。本システムでは，既存の同義語類似語知識ベースによりキーワードの拡張をする。

〈5・2〉 前処理過程

質問回答データベースと文書データベースの更新に対応するために，索引語の抽出と索引語データベースへの更新を前処理部で自動的に行う。また，文書（文）中に現れる索引語を高速に検索するための索引語・文書行列も自動的に計算し，データベースの更新を動的に維持する。これにより，新しい情報に応じデータベースの自動的な更新が可能になる。

索引語の抽出では，キーワード抽出と同じように，まず形態素解析により付与された 39 種類の品詞タグを分析し，その中から助詞，形容詞，及び一般的な副詞と動詞などの意味内容を直接表現しない単語(機能語)と挨拶の言葉を除く。残った品詞には一般名詞(n)，人名（nr)，地名(ns)，機構組織名(nt)，他の固有名詞(nz)，普通動詞(v)，名動詞(vn)，時間詞(t)，方位詞(f)，場所詞(s)が含まれる。次に残った単語の中から不要語データベースに登録されている不要語を除去して索引語を抽出する。本シ

ステムでは，検索は質問応答データベース中の類似質問検索，類似文書検索と類似度の高い文書からの類似回答文検索という三つの検索プロセスからなるので，これらに対応する索引語抽出プロセスも三つの存在する。

〈5･3〉　検索プロセス

システムはまずユーザ質問文と質問回答データベース中の各質問文の間の類似度を計算し，類似度の高い質問文に対応する回答文を検索結果とし応答を返す。これによって，検索結果が得られなかった場合，もしくはユーザが回答に満足できなかった場合は，旅行情報に関する文書の検索を行う。このとき，ユーザ質問文と文書中の文との類似度も計算し，直接質問文に類似した適当な文を回答にする。本検索中で一番重要なことは質問文と検索文（質問回答データベース中の質問文と文書中の文の総称）の類似度を計算することであるが，本システムでは現在の主流であるベクトル空間モデルを用いる。

回答候補の絞り込み：本システムの基本的な考え方は，二種類のデータベースの検索を統合する質問応答システムである。まず，ユーザの質問文と質問応答データベース中の質問文の類似照応を行う。旅行質問において類似質問の集中度が比較的高いという特性により，典型的な質問と過去の重要な質問をあらかじめ質問回答データベースに登録した。検索時は，検索質問文とこのデータベース中の質問文との類似度を計算し，類似度の高い上位三文の回答を出力する。将来的には閾値を用いることも検討しているが，現在，次に示す文書検索を行うかどうかの判断はユーザに任せている。具体的には質問回答データベースによる回答にユーザが満足できない場合に，文章検索ボタンをユーザがクリックする設計になっている。

次には，旅行情報文書データベースの検索。旅行者の様々なニーズに応えるために，質問回答データベースで回答を得られない場合，或いはユーザが回答を不十分だと思う場合には，対象情報源として大量の旅行情報に関する文書を検索する。まず，質問文に類似した文書を旅行情報文書データベースから検索し，類似度上位 3 文書を得る。次に検索された文書中からユーザの質問文との類似度が大きい上位三文を選ぶ。また，一般的に，回答が類似度の高い文の前後にある可能性も高いので，本システムでは，正確率を上げるため，検索された上位三件文の前後の質問文の任意キーワードを含ている文も含め，最終の回答文として回答する。また，本システムには質問応答データベースの動的な更新と疑問詞情報により予め用意したテンプレートを用い聞き返し文の作成もできる。

6. 実験と評価

評価実験のために，2 種類のデータを収集した。一つは質問応答データベースの作成用データで，他方は旅行情報文書データーベース用データである。我々はインターネットで日本旅行問合せに関する中国語のウェッブサイトから 730 の質問応答レコードと約 15000 文を含むハイパーテキスト(HTML)ファイルを収集した。

実験用質問の収集について，我々は質問のアンケートを行い，日本への観光についての 115 の質問文を実験用の質問文として回答を返す実験を行った。

実験の結果，115 の質問に対し質問応答データベースから類似度の上位 3 つで，正しい回答が得られた割合は 26.8%であったことから，質問データベースだけでの検索の不充分性も示している。文書に回答検索実験の結果について，自然言語による質問に対し本論文で提案した方法により類似度の値で回答文をソートして上位三位までに正解があれば正解とする場合には，80%を超える正確率が得られた。質問文の長さと正解率について詳しく見てみると，115 質問文の中で最短は 6 文字で，最長は 25 文字であった。このうち 14 文字以下の質問文が約 78.3%占め，正解率は 83.3%であり，それ以上の質問文が約 21.7%占め，正解率は 76.0%であった。これは長い文では浅い言語解析だけではなく構文等の情報が必要であるとの可能性を示唆している。

7. むすび

本論文では，浅い言語解析による質問文分析及び索引語抽出と VSM モデルに基づき，特定ドメインを向いて，よく聞かれる質問とその回答をデータベース化した質問応答データベースの回答検索と，旅行情報に関する文書の回答検索を統合する QA システムの構築を提案した。インタフェースにおいて従来のテキスト入出力の上に，中国語の文型の特徴を基づき，FSN を用い高頻出定型表現を人手で抽出し観光ドメインへの有限状態文法を構成したうえで，特定領域向きの音声入出力も検討した。

謝辞　本研究は科学研究費補助金基盤研究Ｂ（No. 14380166）によって実施された。

参考文献

（1） Xiaoyan Li. and W. Bruce Croft: “Evaluating Question Answering Techniques in Chinese. Proceeding of HLT 2001”, San Diego, CA. March 18-21(2001)
（2） Steve Young, Dan Kershaw, Julian Odell, et. al, ”The HTK Book (for HTK version 3.0) ” , July (2000)
（3） Huaping Zhang, Hongkui Yu, Deyi Xiong and Qun Liu: “HHMM-based Chinese Lexical Analyzer ICTCLAS”, proceedings of 2nd SigHan Workshop, pp.184-187(2003)
（4） http://mtgroup.ict.ac.cn/~zhp/ICTCLAS/documents.html
（5） Zhengdong Dong et al: Hownet. http://www.keenage.com
（6） Salton: “Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer[M]”, Addison-Westey , Reading, Mass(1989)
（7） 黒橋禎夫：「大規模テキスト知識ベースに基づく自動質問応答」，情処研報，音声言語処理研究会報告書，2001-SLP-39-25 (2001)