ウィキペディアを活用した仮想人物応答システムの研究

神保　由佳里　　(指導教官: 増永 良文)

1　はじめに

古代の遺跡や都市を仮想世界で再現する技術が進んでいる. そこには人間が住んでいたはずだが, それらは再現されていない. 人が再現された場合, ユーザはその人に様々なことを問いかけてみたくなる. ここで問題となることは, 仮想世界で現れる人物のデータをいかにして取得するかである.

現在, インターネット上で自由に誰でも利用できる百科事典であるウィキペディア [1] のデータベースは, ダウンロードして再利用することができる. そこで本研究では, ウィキペディアのデータを活用して, 過去に存在した人物をコンピュータ上で仮想的に作り上げて対話することを目的に, 仮想人物応答システムを構築することとした.

2　先行研究とその問題点

先行研究 [2][3] にて古和らは, インターネット上のウェブコンテンツによる歴史上の人物の情報を活用して, その歴史上の人物を仮想的に作り上げ, 事実に関する単純な質問に対して対話できるシステムの実装を行った. 以下に要点を示す.

1. 時間に関する質問を音声で投げかける.
2. 音声認識ソフトで認識された質問文からキーワードを抽出し, そのキーワードの類義語も交えて検索式を生成.
3. GoogleWebAPI を使用して, 該当人物関連ページの Google 検索上位 10 件のウェブページをファイルとして自動ダウンロードし, HTML タグを除去し, プレーンテキスト化する.
4. 質問より得たキーワード及び類義語で全文検索することによって, 回答候補を抜き出す.
5. 回答候補を集めランキング化し, 最上位のものを回答として対話インタフェースに返し, 音声合成ソフトを用いて発話する.

先行研究では事実に関する単純質問（時間の質問のみ) に対する実装ができている. しかし, まだその他の単純質問（〜はどこか？, 〜は何か？）には対応できていない. また自動でウェブページを取得できるものの, 回答の正答率が低く, かつ回答時間に平均 1 分を要するなどの問題点があった.

実際, 個々のウェブページは文字コードも形式も違い, またノイズデータも多い. ゆえに, Google を用いてウェブ全体を検索の対象としていることが, 正答率の低さに繋がっているのではないかと考えた. そこでネット上で利用できる百科事典であるウィキペディアのデータデースに着眼した. ウィキペディアなら, 人物に関する記事は約 7000 本以上ある. 更にスタイルマニュアルが定められているので, ウィキペディアの記事には大まかな構造上の一貫性が維持されている. よって, ウィキペディアのデータベースを予めダウンロードし, そのデータベースのみを検索の対象とすることにより, 先行研究の問題点に対処できると考えた.

3　対話システム概要

仮想人物に対する質問に関しては, 先行研究と同様に単純な質問のみとする.

Research on a Virtual Person Answering System utilizing the Wikipedia by Yukari JINBO

また, 人物に対する質問は, 「小説家」,「政治家」といったカテゴリごとに対して問いかける質問はそれぞれ共通する質問が多い. そこで本研究では質問対象とする人物を「明治から昭和にかけて活躍した小説家」とすることにした. そこでウィキペディアの小説家である人物記事に何が記載されているか調査した結果, 誕生日, 没年 , 出版日といった「時間」について, 生誕地, 留学先といった「場所」について, その他に本名, 仕事, 出来事といった「イベント」について多く記述されていたので, 「時間」,「場所」,「イベント」については仮想人物が応答してくれるのではないかと考え, これらの質問に対する実装を行った. ウィキペディアを活用した仮想人物応答システムの概念を図 1 に示す.

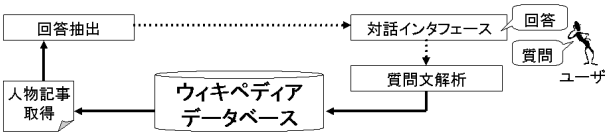

図 1: ウィキペディアを活用した仮想人物応答システム概念図

3.1　対話インタフェース

インタフェース画面を図 2 に示す. ユーザの発話による質問が入力され, 回答抽出モジュールから返ってきた回答を音声出力変換ソフトで読み上げる.

図 2: 対話インタフェース画面

3.2　質問文解析

対話のやり取りは自然言語で行われるが, 現在のコンピュータが自然言語の意味を理解することは難しい. そこで質問文を質問タイプ (「時間」,「場所」,「イベント」) に分類する. この際正確に分類するために, 時間は「〜はいつですか？」, 場所は「〜はどこですか？」, イベントは「〜は何ですか？」と質問形式を限定させ, これらのフレーズを手がかりに分類する. そして, 質問文の「〜」部分に当たる語句を抽出することで質問文から検索キーワードを抽出する. 例えば, 「留学先はどこですか？」という質問から「留学先」を検索キーワードとして抽出する. 本研究では, 音声入力・認識のため使用する AmiVoice SDK において, 音声認識する言い方を限定し, 言い方すべてを網羅的にデータ化するセンテンスパターン認識方式をサポートしているので, このア

プローチは実装可能である.

また, 検索キーワードの類義語は予め定義しておく. そして検索キーワードと質問タイプから何を問う質問か判定し, その回答を抽出するプログラムに飛ぶ. 例えば「誕生」に対して「生誕」,「生年」,「出生」,「生まれ」を類義語として予め定義する. そして検索キーワードが「誕生」に関する類義語で質問タイプが「時間」なら, 質問文解析の結果, 誕生日を尋ねる質問とみなし, 誕生日を抽出するプログラムに飛ぶ.

3.3 人物記事取得

ウィキペディアデータベースから該当人物のデータをテキストファイルとして取得する. なおこのテキストファイルは, ウィキペディアのマークアップが付いたままになっている. マークアップとは, ウィキペディア専用の HTML タグのようなものである. 例えば, 「'''夏目漱石'''」のように「'''」で囲むことにより,「**夏目漱石**」と太字で表示される.

3.4 回答抽出

先行研究では, 複数のウェブから抽出した回答候補をランキング化し, 最上位のものを回答としていた. しかし, 本研究では回答抽出の対象がウィキペディアの記事のみである. そのため, 質問内容ごとに回答抽出方法を変えることにより, 回答が高い正答率を示せるように考慮した. 回答抽出の方法は 3 パターンに分かれる. 以下にそれぞれの詳細について述べる.

3.4.1 スタイルマニュアルを活用できる質問

スタイルマニュアルを活用できる質問としては, 生年月日, 没年月日, 仕事を問う質問が該当する. スタイルマニュアルとは, 「記事を編集する際, 全体として守らなければならないこと」である. 人物に関するスタイルマニュアルも存在し, そこには人物記事の第一文には「人物名（振り仮名, 生年月日ー没年月日）職業.」と記述するように指示されている. この部分を抽出すれば, 生年月日, 没年月日, 仕事に関しては正しい答えを抽出することができる.

3.4.2 年表などを活用できる質問

年譜, 受賞歴, 作品一覧といった年代順に箇条書きで書かれている部分を活用できる質問としては出版日, 出来事を問う質問が該当する. 年表などは, 個々の人物の重要な出来事について年代順に箇条書きで書かれているので, 人物記事の本文にありがちな「三年後」などといった言葉がないため日付補正をする必要がない. よって質問内容に対して, 正確な年や出来事を抽出しやすい. 尚, ウィキペディアのマークアップにおいて「＊」を行頭に書けば, 箇条書きとなる. 年表では「＊」ごとにセンテンスとみなし, 分析する. 例えば「羅生門を出版したのはいつですか?」という質問なら, まず人物記事から年表を抽出し, そこから「羅生門」というキーワードを含むセンテンスを見つけ, 日付表現を抽出する.

3.4.3 全文検索で回答抽出する質問

上記の質問のように検索する場所を限定すると回答を抽出できない可能性がある質問としては, 留学先, 生誕地, 本名を問う質問が該当する. これらの質問は, 人物記事に対して全文検索を行う. これらの回答を抽出する際, 「回答候補のうち本当の回答は, 統計的に質問文中のキーワードと近い位置にある」という方針に基づいて行う. また助詞に注目することで正確な解答を抽出できるよう考慮した. 例えば「留学先はどこですか？」という質問に対しては, まずテキスト中の文章を句点ごとにセンテンスとみなし分析し, 検索キーワード (またはその類義語) である「留学」を含むセンテンスを抽出する. 抽出されたセンテンスに対して, 方向に続く助詞である「へ」や場所に続く助詞である「に」を検索し, その直前のセンテンスを回答として抽出する.

4 実装環境

以下の仕様で対話システムを構築する

- OS:WindowsXP Home Edition
- 音声入力ソフト:AmiVoice SDK 5
- データベースシステム:MySQL Server 5.0.15
- データ検索:Perl
- 対話インタフェース:VisualBasic
- 音声出力変換ソフト:NEC Smart Voice 6.0

尚, ウィキペディアのデータベースをダウンロードするにあたり, MediaWiki 1.8.2 というウィキペディアで用いられているソフトウェアと Apache HTTP Server 2.0 というウェブサーバソフトを利用し, ウェブ上で提供されているウィキペディアのデータ (314.8MB) を MySQL へインポートした. データは, page(page_id,…, page_title,…), revision(…, rev_page, rev_text_id,…), text(old_id,…, old_text,…) という三枚のテーブルに格納する. これにより, ウィキペディアのサイトに接続しなくても, データベースから自由にウィキペディアの記事を取得することが可能となる. また, 本研究で使用した明治から昭和にかけて活躍した小説家 9 名分の記事は, 合計 129KB である.

5 実装結果

構築したシステムを使って, 明治から昭和にかけて活躍した小説家を任意で選び, 実験を行った. 本研究で質問可能となった質問は, 生誕日・没年日・出版日・生誕地・留学先・仕事・本名・出来事を問う質問だが, いずれもほぼ 100 %近くの回答率を示した. また, 先行研究では回答に対して 1 分近くかかっていたが, 本研究では 3 秒以内に回答が抽出でき, リアルタイムな回答を得ることができた.

6 まとめと今後の課題

本研究での対話インタフェースは先行研究をベースにした. 本研究では, まずウィキペディアのデータベースをダウンロードした. そして, 人物に対して単純な時, 場所, イベントに関する質問を投げかけたら, 質問文解析をして, 対話したい人物ついての記事を取得し, そこから回答抽出をする Perl のプログラムを新規開発した. その結果, 先行研究の課題のうちの「場所」,「イベント」に関する質問への対応, 回答時間の短縮, 並びに回答内容の正答率の向上を達成することができた.

本研究では個々の質問に対して, 回答抽出するプログラムを組んでいる. そのため高い正答率を得ることができたが, 現段階では個々の人物記事の 3 割程度の情報しか利用できていない. そのため, センテンスパターンに対応していない質問をされた場合の対処が難しい. 今後の課題としては, 質問文のセンテンスパターンを増やし, ウィキペディアの記事を活用してより多くの質問に対して, 自動で回答を抽出していくことが必要である.

参考文献

[1] “フリー百科事典ウィキペディア” http://ja.wikipedia.org/wiki/

[2] 古和美由紀: “ウェブコンテンツを活用した仮想人物応答システムの研究,” お茶の水女子大学理学部情報科学科第 12 回卒業研究発表会要旨集, pp.33-34, 2005.

[3] 郡司京子: “ウェブコンテンツを活用した仮想人物応答システムの研究,” お茶の水女子大学理学部情報科学科第 13 回卒業研究発表会要旨集, pp.71–72, 2006.