百科事典を対象とした質問応答システムの開発

関根 聡
ランゲージ・クラフト研究所*
sekine@languagecraft.com

1 序文

質問応答システムの技術が注目されている。これは、物事に関して尋ねる質問を自然言語文で入力し、その答そのものが返ってくるという技術である。この視点はかなり古くからあったものであるが [Winograd 1977], [Woods and Kaplan 1977]、最近の研究は新聞記事などを対象にした情報検索の発展形として提案されたものであり、情報抽出の技術に端を発した固有表現抽出の技術の成熟とあいまってかなり高いレベルの正解率を出すようになってきている。新聞記事を対象とした評価型のプロジェクトが、米国では TREC の QA タスクとして [TREC QA Homepage]、日本では NTCIR の QAC タスクとして [NTCIR QAC Homepage] 取り挙げられた。それらのプロジェクトは数多くの参加者を得、参加者は基礎的な技術の蓄積と、新しいパラダイムとしての応用の可能性を模索している。これらの評価型プロジェクトでは、物事 (体言で表現されるもの) を対象に新聞記事から解答を取り出すという前提が置かれている。つまり、解答は新聞記事中にあることが保証されているものが多い[1]。しかしながら、新聞記事に載っている物事は限りがある。例えば、10 年分の新聞記事に、アメリカの歴代大統領の名前がすべて載っているという保証はない。また、根本的に、新聞は物事の説明をすることが目的で書かれているものではなく、出来事を伝達するための媒体である。したがって、物事についてを聞くために新聞を知識源にすることは、なんらかの無理があると考えられる。つまり、一般的な応用を考える場合には、このような設定では本質的に答えられない質問がどうしても存在するであろうと予測できる。(したがって、逆に言うと新聞記事を知識源とするのならば、その新聞記事の範囲内で起った出来事に関する質問をするのが適切だと著者は考える。) では、物事を聞くために、何を知識源とすれば良いだろうか。まず思いつくのが百科事典である。百科事典は、基本的に物事を説明する媒体である。時事的な物事を除けば[2]、人が広く興味を持つような物事を説明することが目的であり、物事を聞く質問応答システムの知識源として最適であると考えられる。

ここでは、このような考えを元に実現された、百科事典を対象とした質問応答システムについて述べる。本論文で紹介するシステムはすでにウエッブで公開されている[3]。小学館の日本大百科全書 (ニッポニカ) を知識源としており、その全体を対象としたシステムは有料会員のみであるが、(ただし、会員ではなくても 2 日間「お試し利用」の制度を用いてシステムを利用することができる) 会員でない人にも、「温泉」をテーマにした「機能限定版」を使用することができる。

2 日本大百科全書の特徴

新聞記事と比較して、日本大百科全書 (以下、全書と呼ぶ) には質問応答システムに利用できると考えられる以下のような特徴がある。

1. 情報の構造化
   見出し語 (項目) と説明文という対応がきちんと取れており、各見出し語は、大分類、中分類、小分類で分類されている。
2. 必須の情報
   また、人名の場合には、生年、没年などが必ず記載されている。
3. 語の定義
   見出し語の定義が、ほとんどの場合に 1、2 文目に書かれており、その後に細かい定義、または背景や状況の説明の文章が続くという説明文の構造がある。
4. 言い換え、類義語、関連語
   言い換え、類義語、関連語などが明示的に示されており、その語の情報も利用することが可能である。

今回の質問応答システムの開発では、このような特徴を利用した。以下のシステム構成で詳しく述べる。

*本会社は関根の個人的なコンサルタント、および、ソフトウェア開発等の受け皿として設立した日本の株式会社である。(www.languagecraft.com) ベンチャー企業として利益を上げるといった事を目的に設立したものではない。関根の主なる所属と肩書は、依然、ニューヨーク大学コンピュータサイエンス学科の研究助教授 (sekine@cs.nyu.edu) である。

[1] 一部には、解答が新聞記事にない質問をある割合で混ぜたり、新聞記事に解答があるかどうかは分らないまま質問を設定し、後で確認するという方法で質問が作成されている場合もある。

[2] しかしながら、今回使用させていただいた日本大百科全書は、年に何回も更新が行なわれており、時事的な物事も記載されている場合が多い。

[3] アドレスは www.japanknowledge.com

3 システム

一般に公開するシステムを目標としたため、評価プロジェクトのような綺麗な質問ばかりとは限らない。特に、質問応答という概念を無視して、キーワード検索を行なったり、「日本の城について。」と普通には質問とは考えられないような質問も入力される。また、関連語へのリンクや見出し語との一致といった方法での検索も実現すると便利である。さらに、ユーザーの入力には、カタカナなどの表現の揺れが存在することも考えられる。全体のシステムは、このような色々な検索方法や仕組が融合した形となっており、ユーザーの入力から適切なサブシステムを選び、そのシステムの結果を適切に返すというものになっている。

基本的な質問応答システムの構成は、新聞記事での一般的なシステムとほぼ同様である [Sekine et al. 2002b]。まず、質問文を解析し、キーワードと、解答のタイプを抽出する。ここでタイプとは、人名、会社名、場所数などの概念的な分類であり、関根ら [Sekine et al. 2002a] が定義したものに対し、百科事典の見出し語を参考に拡張したものを用いた。そして、キーワードを基に、説明文を対象として検索を行なう。ここでは、全書の特徴で述べたように、説明文の 1、2 文目にその見出し語の定義が述べられていることが多いというヒューリスティックから、その範囲にキーワードが表われている場合には重みを増している。もし解答のタイプが数値表現ではない時には、解答のタイプと見出し語のタイプの一致度を検索スコアに加味し、スコアの高いものから順に並べて終了する。解答のタイプが数値表現の場合には、検索された項目の内の上位 10 個に対して、その説明文中から、キーワードとの距離と、タイプの一致度により、解答となる単語を選択して終了する。

今回の開発は、短期間であったため、解答となるのは、すべての見出し語か説明文中の数値表現に限られている。つまり、説明文中にある数値表現以外の表現は解答として抽出されない。したがって、「光源氏の長男の名前は」という質問にも、「夕霧」という項目がないかぎり答は出せない。(実際には、夕霧は「源氏物語の登場人物」や「源氏物語のあらすじ」に登場するのみである。) 数値表現に関しては、その自動的な抽出が比較的容易なことから説明文中であっても解答の対象とした。

このような質問応答システムの他に、質問に答える仕組を用意してある。それは、いくつかのよくある決ったタイプの質問に対しては、典型質問パターンを用意し、そのパターンにマッチした場合には本来の質問応答システムによる解答の探索は行なわずに、直接答を返すという仕組である。例えば、生年を尋ねる質問である。「川端康成が生れたのはいつですか？」というような、誰かの誕生日を聞く質問は、驚くことに、大きく分類すると 3 種類、少し細かく分類すると 29 種類のパターンで、ある程度考えられる質問をカバーするパターンを作成することができた。このパターンにマッチすると、全書の特徴で述べたように、全書では人名の生年、没年が特別に記してあるので、その人の記載があれば、その人の生年をそのまま返すような仕組を導入してある。

キーワード検索はスペースを置いた AND 検索のみで、論理式には対応していない。カタカナなどの表現の揺れは、見出し語にない入力があった場合には似た表現を探し、それを表示する仕組を取り入れている。たとえば、「コラアゲン」と入力しても「コラーゲン」が探し出される。また、全書の特徴で述べたように、言い換え表現も全書に明示的に示されている。例えば「ギョーテ」は「ゲーテ」に、「リューマチ」は「リウマチ」にリンクされている。このような表現の揺れに対処することによって、ユーザーに優しいインターフェースになっていると考える。この機能はキーワード検索以外の部分にも拡張していく予定である。(一部はすでに、質問応答システムでも実現されている。)

知識の作成、整理については、全書にもともとある分類を参考にしながら、我々の固有表現階層のタグをすべての見出し語に人手で付与した。説明文中の数値表現については、ある程度自動で行ない、タグ付けされなかった数値表現については、頻度の高いものや簡単にパターンになるような物を対象に人手でタグ付けを行なった。

質問応答システムの初期の検索画面の例を図 1、結果の画面の例を図 2 に示す。初期画面においては、「どんなことを調べたいですか」とある下のボックスに、質問を入力し、「教えて!」のボタンを押すと、システムが動作する。普通、システムを動作させると、大抵の質問に対して 0.1-0.5 秒程度で結果画面が表示される。

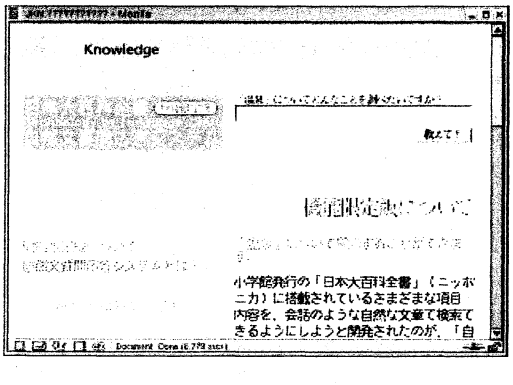

図 1: 検索画面の例

本システムは、会員専用版も機能限定版も、質問者のログを記録している。新たに入力された質問例は、一般のユーザーからの自然な質問として貴重なデータである。今後のシステムの精度向上や機能拡張のための、他にはないデータである。基本的には、このデータは研究

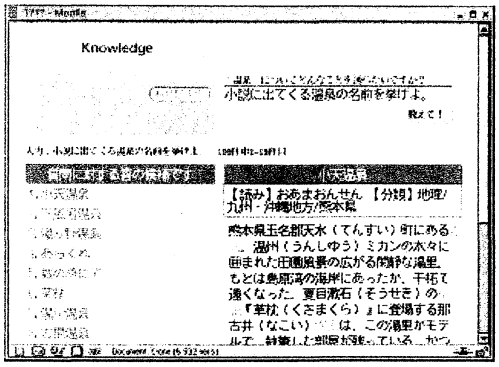

図 2: 検索結果画面の例

目的で公開できるように交渉中であり、質問応答システムの研究を行なう研究者に配布できることを目標にしている。

4 評価

システムの評価結果を述べる。評価は、多くの学生、主婦のアルバイトの方に作成していただいた 500 質問例の内の、解答が百科事典の中にない 135 質問を除いた 365 質問で行なった。結果を表 1 に載せる。

名詞表現では、正解が項目に上っているか、説明文中にあるかのいづれかのものが合計 302 例あった。その内、今回のシステムが対象とするのは項目に挙がっている 175 例である。20 位以内にその項目を見付けた数は 115 例であり正解率は 65.7%であった。正解が説明文中にあるものも含めると、302 例中の 244 例となり、正解率は 80.8%になる。

数値表現については、1 例を除いて、すべて説明文中にある。正解を含む項目を出せたのが、63 例中 47 例 (74.6%)、正解を正しく出せたのが 28 例 (44.4%) であった。

総合すると、239 例 (175+1+63) について正解を出せる内、システムが正解を出せたのが 144 例 (144/(115+1+28) =56.2%) であった。また、なんらかの形で正解を見付けられるべき質問の数は 365 例あり、その内の 291 例については、なんらかの形式で正解を見付けられ、その割合は 79.7%であった。

5 新聞記事システムとの比較

知識源の違いによる正解率の比較をするために、上記の評価で用いられた 500 質問文をＮＹＵ／ＣＲＬで開発された新聞記事を対象とした質問応答システム[4]に入力し、その結果を比較した。結果を表 2 に載せる[5]。百

百科事典＼新聞記事	正解	正解を探せない
正解を出した	58	87
項目を見付けた	41	104
正解を探せない	15	59
正解がない	12	124
合計	126	374

表 2: 比較結果

科事典の場合には、項目の内の説明文から解答を抽出しているのは数値表現のみであり、本文から正解を探している新聞記事のシステムと直接の比較は容易ではない。つまり、百科事典が完全な正解を出せた場合のみを見ると、百科事典が 145 であり、新聞記事の 126 と百科事典の方が多少多い程度であるが、項目の中にも正解がある場合を換算すると、百科事典の方が正解を出せる可能性はかなり高いことが分る。具体的な分析をしてみた。百科事典でできて、新聞記事では正解を探せなかった質問では、40 の質問について、答を利用したキーワードサーチなどで細かく新聞を分析したが、新聞には答がないと判断されるものが 26 例であった。(例えば、「音の速さは」「アテネのアクロポリスにある神殿の名前」) この割合を全体に適応すると、新聞記事に正解がないと予想される質問数は 243 質問となり、500 質問中、約半数を占める[6]。残りは、表記の揺れ、その他、システムの未熟性によるものであった。逆に百科事典で見付けられず、新聞記事で正解を探せた質問では、表記の揺れ、同義語が原因なのが 1/3 (例えば、レオナルド・ダビンチ - レオナル・ド・ダビンチ、向日葵 - ヒマワリ、慶應大学 - 慶應義塾大学)、簡単な言い換えなどを用い質問の表現を少し換えると正解が出せるのが 1/3 (平安京を起した - 平安京に遷都した)、推論が必要なのが 1/3 程度 (1 年は何の日で終る - 1 年の最後の日は) であった。百科事典には正解が存在せず、新聞記事のみに正解があるものは、12 とかなり少ない。質問から正解を導く際に推論が必要な場合にも、たとえその正解があっても正解は存在しないとして判断した (例えば、「関羽の義兄弟は何人」という質問に対し、百科事典に「全部で 3 人の義兄弟」から「2 人」という答を出すには推論がいる)。このようなものを除くと、12 例中で本当に正解が見付からない質問は、「人間失格を書いた作者の絶筆は」のみであった。両方とも正解がないと思われる質問には、「冷蔵庫はどこに置く」「魚はどこに住む」「赤信号の次

[4]本システムは、NTCIR-QAC1 の評価では、知識源の 98、99 年度の新聞記事を予め人手で解析し、システム構築に利用していた上位 2 システムを除くと、2 番目の成績であった。

[5]表 1 と 2 では若干のデータの違いがある。これは、新聞記事を見て正解を変えたものであるが本質的な内容には変化はない

[6]実際は両方できないものはより難しいので、300 質問近くは正解がないものと予想できる

	名詞表現	数値表現	合計
質問応答例	302	63	365
正解が項目に挙がっている	175	1	176
正解が説明文中にある	245	63	308
正解の項目を見付けた	115	1	116
正解を説明文中で見付けた	0	28	28
正解を含む項目を見付けた	212	47	259
なんらかの形で正解を見付けた	244	47	291

表 1: 評価結果

は何」といった常識や当たり前の質問のほか、「お米を計る単位は」「一時間は何分」といった単位に関するもの、漠然とした質問「日本と同程度の食料自給率の国」「猿は何を嫌う」や、逆に専門的なもの「バーコードの日本の国番号」「ショパンの幻想即興曲は何年に作られた」といった色々な種類の質問があった。

これらの結果から、我々が集めた物事を尋ねる質問に対しては、百科事典の方がより適していると言えそうである。

6 まとめ

小学館の百科事典、日本大百科全書 (ニッポニカ) を対象とした質問応答システムの開発について述べた。質問応答は、自然言語処理の技術の中でもかなり実用に近いと考えられる。応用の場を模索したり、視点を変えると、様々な形で世の中に出ていけるのではないだろうか。そして、実際に使われていくことによって、自然言語の研究のあり方や価値が広く問われ、この分野の研究の健全な発展に寄与するものと考えられる。本来、この開発は、著者が百科事典に対する質問応答システムの研究をするために、百科事典の使用許可をいただきたい、と小学館に伺ったところから始まった。小学館とネットアドバンスの方々の理解と協力のお陰で現段階まで到達することができたのは幸運であったが、その他の我々の研究においても、このような場が隠れている可能性があるのではないだろうか。これからも、このような視点も忘れずに研究を続けていきたいと考えている。

7 謝辞

本プロジェクトは、多くの人の協力によって実現された。まずは、以前から百科事典に対する質問応答システムの実現を求め、現在の技術レベルを理解した上で、その実現に協力していただいた株式会社ネットアドバンスの方々に感謝する。特に、鈴木様、相原様には様々な面で協力をいただいた。また、私の無鉄砲な行動を暖かく見守り続けていただいているニューヨーク大学のGrishman 教授には改めて感謝を述べたい。技術的な開発においては特に以下の方々の直接、間接的な協力をいただいた。ここに名前を挙げさせていただき感謝の意を表わす (順不同、敬称略)。須藤、新山 (NYU)、野畑周、内元、井佐原 (CRL)、宮口、脇寺、桝井 (三重大学)、安藤、伊藤、岡田、小川、渡邊、石崎 (慶應 SFC)、河田康弘 (エセックス大学)、岡村、河田篤子、栗田、竹内、野畑恵理子 (フリー)。

参考文献

[Sekine et al. 2002a] Satoshi Sekine, Kiyoshi Sudo, Chikashi Nobata, "Extended Named Entity Hierarchy" *In the proceedings of LREC 2002*, 2002

[Sekine et al. 2002b] Satoshi Sekine, Kiyoshi Sudo, Yusuke Shinyama, Chikashi Nobata, Kiyotaka Uchimoto and Hitoshi Isahara "NYU/CRL QA system, QAC question analysis and CRL QA data" *In the proceedings of NTCIR workshop 3 - QAC1*, 2002

[Winograd 1977] T. Winograd, Five lectures on artificial intelligence. Linguistic Structures Processing, In *Fundamental Studies in Computer Science*, 5:399-520, 1977

[Woods and Kaplan 1977] W. Woods and R. Kaplan, Lunar rocks in natural English: Explorations in natural language question answering, In *Fundamental Studies in Computer Science* 5:521-569, 1977

[TREC QA Homepage]
http://trec.nist.gov/data/qa.html

[NTCIR QAC Homepage]
http://www.nlp.cs.ritsumei.ac.jp/qac/