属性値の自動抽出と質問文パターンを使った
百科事典質問応答システム

関根聡*1　　須藤清*1　　安藤まや*2
*1株式会社ランゲージ・クラフト研究所
*2慶応義塾大学SFC

１．目的

本論文では、百科事典の各項目の説明文中にある属性値を対象とした質問応答システムについて述べる。例えば、「イグアナの全長はどのくらいですか」という質問に対して答えたいとする。百科事典の「イグアナ」という項目の説明文を読むと「イグアナの体長は１．５～１．８メートルで…」と書いてある。本論文で提案するシステムでは、ここに書いてある情報を、「項目：属性：属性値」のタプルの形で「イグアナ：体長：１．５～１．８メートル」として予め抽出しておき知識として溜めておく。そして、質問応答システムでは、ユーザーから入力された質問文から、この質問文ではこの項目のこの属性を聞いているのだな、ということを判断し、抽出されている属性値があればそれを返す。この手法は、現在QACなどで主流の、テキストから動的に答えを探し出すという方法と違い、以下のような問題を解決しなければならない。

１）属性を設定する
２）百科事典の説明文から属性値を抽出する
３）質問文からどの項目のどの属性値を質問されているかを判断する

各問題に対して、我々のとったアプローチを３章以降で紹介し、最後にシステムの評価結果を報告する。

２．背景

現在、主に研究されている質問応答システムの手法は、１）質問文を解析し、キーワードと答えのタイプを判定する。２）抽出されたキーワードを使い情報検索により答えのありそうなテキストを検索する。３）検索されたテキストから期待されたタイプの答えを見つけ出す。という３段階で実現されている（TREC）(QAC) (関根 2003)。しかしながら、キーワードの抽出や拡張の不安定さ、タイプを判定する固有表現抽出の精度の問題などから、比較的に簡単な問題でもきちんと答えが出せないケースがある。また、質問応答システムに入力される質問を分析すると、それにはいくつかの種類があることがわかる。したがって、ある特定の質問タイプに対して、そのタイプ固有のより精度の高い方法があれば、その方法で答えを導くことが有効である。特に、「動物の全長」「人の誕生日」のような「ある事柄の属性値」を尋ねる質問は、質問応答システムでよく聞かれるタイプの質問である（高橋他 2004）。本研究が対象にしている百科事典では、このような属性値の記述が頻繁に見受けられ、情報を比較的綺麗に構造化しやすく書かれているため、属性値の自動抽出を高い精度で実現でき、既存の方法とは異なった質問応答システムを実現できると考えられる。

本手法の関連研究として（Ravichandran et al. 2002）がある。彼らは、まず、人名と誕生日のような典型的な関係を示す表現を、「モーツアルト」と「１７５６」というキーワードでテキストを検索することによって抽出する。そして、その表現を使ってWEBなどから予め人名と誕生日の知識を獲得し質問応答システムに使うことを提案している。ただし、その研究で対象とされている関係の種類は数種類に限られている。また、（高橋ら 04）は新聞記事テキストから属性属性値の情報を抽出する方法を試みている。新聞記事にはそもそも属性関係が少なく，その種類もかなり限られているが、このような情報を自動的に抽出しようという研究は興味深い。その他に関連する研究としては、(Brin 98) (Agichtein and Gravano 00) (Lin and Pantel 01)　(Fleischman 03) が挙げられる。しかし、我々が本論文で提案しているように多くの属性を対象としたもの、百科事典を対象としたもの、質問文パターンをここであるような方法で大規模に作成したもの、一般的な質問応答システムと統合的に作成しているものは筆者の知る限り存在しない。

３．属性の設定

今回の質問応答システムにおいて対象にする属性の設定方法について説明する。対象にしている百科事典は全部で約11万項目あるため、全ての項目に対して人手で属性を決定するのは現実的ではない。幸いにも、我々は過去に行った百科事典を対象とした質問応答シ

ステム（関根 2003）で、全項目に対して意味的な分類を行った。固有表現とは人名、組織名などの名前を中心とした意味的カテゴリで、百科事典の項目では140種類の分類を利用している(Sekine et al. 2002)。各項目の属性はこの分類ごとに似通ったものになると考えられる。そこで、我々は、分類ごとに最低10項目、最大100項目（最も頻度の高い人名）をランダムに取り出し、それらの説明文を読んで、可能な属性を抽出するという作業を行った。そして、基本的に2項目以上の例で共通の属性があったものをそれぞれの固有表現カテゴリの属性として設定した。

４．説明文からの属性値の抽出

次に、各項目の属性値の抽出について説明する。属性値の抽出は、上記で説明した10~100項目に対する属性と属性値のデータを学習データとし、文脈上にある単語を素性とする簡単なルールを、頻度を基準に学習し、それを使って行った。MEやSVMなどのより高度な機械学習システムを用いなかったのは学習データが少なく、それには適さなかったためである。

ルールは百科事典項目の説明文中から抽出する。属性値のコンテキストや属性値そのものにおける単語自身、品詞や固有表現の種類を基にしたパターンであり、そのパターンがトレーニングデータ中でどれだけ現れているかという指標に基づきスコアリングする。また、百科事典では1文目や頭の数文目までに重要な属性値が現れることが多いため、文の位置の制約も付けられるようになっている。パターンは、1つの単語に関する情報（パターン要素）を集めたものからなっており、どのような組み合わせでパターンを作るかは25種類のパターンテンプレートによって限定されている。パターン要素の単語が属性値に対してどの位置にあるかは、４つの情報の組で示される。

1つ目の情報は属性値の持つコンテキストの種類を指定するもので、属性値の外側のコンテキストを素性（外的素性）とするものと属性値そのものの単語･文字列そのものを考慮した内側のコンテキスト（内的素性）の二つに分類し、外的素性を”O”、内的素性を”I”として表す。2番目は、外的素性、内的素性のそれぞれは属性値の語頭の境界を指定する左素性、語尾の境界を指定する右素性と分類され、左素性を”L”、右素性を”R”で表す。例えば、「体長１．５メートルの」という文からは体長という属性に関して「体長」は属性値の外左素性（OL）、「の」は外右素性（OR）として得られることになる。三番目の情報はコンテキストを、何を単位として見るかを示し、W（単語表記そのもの）、B（単語の原形）、P（単語の品詞）、N（単語の固有表現クラス）のいずれかを取る。最後の情報は属性値前後の境界からの単語数を示す。

例えば、「体長/は/１．５/～/１．８/メートル/で」という文からは、属性値前後の境界からの単語数を２以下に絞り単語の情報のみで（OLW2：体長）、（OLW１：は）、（ILW１：１．５）、（ILW２：～）、（IRW２：１．８）、（IRW１：メートル）、（ORW１：で）というパターン要素が抽出される。これらのパターン要素をパターンテンプレートにしたがって組み合わせて、パターンの候補を作る。次に、パターンの候補が学習データ中にいくつ現れ、そのうちのいくつが実際に属性値を示しているかを数える。全頻度がある敷居値以下で、属性値を示している頻度が２以上であるときに、そのパターン候補はルールとして登録される。

以下にパターンテンプレートの一部と「爬虫類の体長」に関して学習されたルールの一部を示す。ルールの頭にあるの”*”はそのパターンが説明文中のどこに現れてもいいことを示す。この値は、先頭文でなければいけないことを示す”1”や3文目までになければいけないことを示す”3”になったりする。

パターンテンプレート：

PATTERN ILW2 ILW1 ORW1 ORW2
PATTERN IRW1 IRW2 IRW3
PATTERN ILN1 ILP2 ILN3
PATTERN OLW2 OLW1

学習されたルール（爬虫類の全長）：

* PATTERN (OLW2:、) (OLW1:全長) (ORW1:。) (ORW2:$) … (3/4)
* PATTERN (IRW1:メートル) (IRW2:２．５) (IRW3:～) … (2/16)
* PATTERN (ILN1:OTHERS) (ILP2:名詞) (ILN3:PHYSICAL_EXTENT) … (2/16)
* PATTERN (OLW2:、) (OLW1:全長) … (5/38)

学習されたルールは属性値抽出の対象となるテキストと照らしあわせた上、ルールに該当する部分があればマッチしたものとされる。ルールに左素性が含まれていれば属性値の語頭の境界が素性中の境界からの距離によって計算されスコアが加算される。同様に、右素性が含まれていれば属性値の語尾の境界としてのスコアが加算される。境界としてのスコアはマッチし得るルールに与えられたスコアの合計として求めている。また、同一文中で複数の境界の候補が出た場合には最もスコアの高いものを取る。

５．質問文パターン

質問応答システムに入力された質問文から、その質問文が何という項目の、どの属性の値を求めているのかを解析するために質問文パターンを作成した。今回対象としているのは項目の属性値を聞く質問であり、そのバリエーション（言い換え）は限られていると考えられる。そこで、我々は属性値を聞く質問を、正規表現で記述できると考え、試みた。百科事典に見つかった約80種類ある項目の固有表現種類のそれぞれに挙げられた属性を合計すると約800ある。しかし、例えば「爬虫類の全長」と「魚類の全長」を聞く質問の構造はほぼ同じであると考えられる。実際に、例えば「橋の長さ」や「川の長さ」など長さを聞く質問の構造もほぼ同じ形式をしている。したがって、このような質問の目的によって属性値の種類をまとめてみた。すると、800の属性の内、名詞句が属性値となっているもので且つ、質問文パターンにできそうな454の属性は、長さ、重さ、開始時間、場所を聞く質問など31種類の質問タイプに分類できた。そして、それぞれのタイプに対して、まず、構文的な質問のパターンを作成する。そこでは、これまでの研究を進める過程で集めた質問文を参照して、構文的な構造が同じようなものを集め質問文パターンを作成していき、その後で、語順を変えるなどしてバリエーションを増やしていった。長さを聞く質問文の構文パターンの一部を以下に挙げる。

（項目）の＜長さ＞はどのくらい
（項目）の＜長さ＞はどれくらいか知りたい
（項目）はどれくらい＜長い＞ですか
どれくらい（項目）は＜長い＞ですか

このようにして、構文的な言い換えに対応する質問文パターンを作成したが、上記のパターンで＜長さ＞の部分は、それぞれの項目の種類と属性の組み合わせごとに特有である。例えば、爬虫類の全長の場合には、「全長、体長、長さ、大きさ」などになり、橋名の橋長の場合には「橋長、全長、長さ、距離」などが入る。このような語彙的知識はそれぞれの項目の種類と属性の組み合わせに対して個々に作成してある。また、文末の「どれくらい」や「知りたい」「ですか」といった表現は全ての属性に対して共通に用意してあり、例えば、文末が「知りたい」が「教えてくれ」や「に関して説明して欲しい」であってもいいように正規表現で用意している。構文パターンは、それぞれの質問タイプにより異なるが、約10～30種類の構文パターンが用意され、＜長さ＞＜長い＞のような変数は、31種類のそれぞれの質問タイプに対して４～８種類の変数として必要であった。属性によって異なるが、シンプルな「長さ」を聞く質問であっても、正規表現を展開すると1万以上の種類の質問（言い換え）に対処できるようになっている。

実際の処理では、質問文のなかで項目の部分が指定されているわけではないので、少し工夫が必要である。本システムでは、質問文中にある部分文字列中に11万項目のいずれかの項目にマッチする文字列があれば、それを（項目）に置き換えて、その項目の全ての属性の質問文パターンとマッチさせる。たとえば、「イグアナの長さはどれくらいですか」という質問だと「イグアナ」と植物の「イ」が項目候補となり、それぞれ、爬虫類の6種類の属性と植物の9種類の属性の質問パターンセットとのマッチングが行われる。

６．評価

開発に関わってない日本人に、百科事典の説明文を読んで質問を作成してもらい、それによって評価を行った。質問は、説明文中にある名詞句で、その項目の属性の値として相応しそうなもの（その項目と同じような種類の項目に頻繁に出てくると考えられる属性）を答えとする質問を300作成してもらった。この300質問をシステムにかけ、どれだけの質問に答えられるか、答えられない場合にはどのコンポーネントで誤っているかを調べた。結果を表２から表４に示す。

表２には属性設定の評価結果が示してある。300質問中、対象になりえない問題として、属性値とはいいがたい物を質問している場合と、今回は何らかの固有表現のタイプが指定されている項目を対象にしていたが、そうではないものの質問であったものが51質問あり、対象になりえる質問は全部で249質問であった。その中で、3章で述べた属性の設定で、属性が設定されているものが43%の106質問であった。この数は半分以下と少ないが、属性の設定時点で、10項目のサンプリングの中だけで属性を決めたこと、そのタイプに共通するものであるという条件をつけたためであると考えられる。属性設定において、いかに幅広い属性を設定するかという問題点が浮かび上がる。

対象になりえる全問題	249
属性が設定されていた	106
属性が設定されていなかった	143

表２：属性設定の評価結果

質問文パターンの評価結果を表３に挙げる。質問文パターンは２つ方針で対象を限っている。一つは、名

詞句が属性値となっているものであること、もう一つは質問文としてパターン化することができそうな属性であることである。その結果、全部で約800ある属性の内、454の属性しか対応していない。この制限のため、106の質問文中29問が質問文パターンの対象外となった。属性が設定されていた106の質問文の中で、この29個を除いた77の質問中、現在開発した質問文パターンでマッチしたのは56個であり、マッチングの割合は73%であった。できなかったものの29質問中には「東京大学の本部の住所を教えてください」のように「の本部」という余計なものが付いた場合や「ルーブル美術館はどこの国にある美術館ですか」というようにタイプを示す名詞句が繰り返されているような場合があった。しかし、29質問中本質的にパターンで対応するのが難しい質問は約10質問あった。また、不適切な属性の質問文パターンとマッチしたものは評価データ中には見当たらなかった。

属性が設定されていた	106
質問文パターンがマッチした	56
質問文パターンがマッチしない	21
質問文パターンの対象外	29

表３：質問文パターンの評価結果

質問文パターンがマッチし、どの項目のどの属性の値を返せばいいか分かった56質問の内、属性値がきちんと抽出できていたものは64%の36質問であった。誤った属性値が抽出されたものが6質問、属性値が抽出されていなかったものが14質問で、少ない学習データを使いシンプルな方法で属性値の抽出を行った割には、高い精度で属性値が抽出できていたことが分かる。学習データを増やした場合に精度がどう変化するか調べたいと考えている。

質問文パターンがマッチした	56
属性値が正しく抽出された	36
誤った属性値が抽出された	6
属性値が抽出されなかった	14

表４：属性値抽出の評価結果

７．まとめ

本論文では百科事典の各項目の説明文中にある属性値を対象とした質問応答システムについて述べた。属性の設定、属性の自動抽出、質問文パターンの利用というアプローチを取った。精度は、それぞれのコンポーネントである程度の精度を得た。今後も各コンポーネントの精度向上を通して全体の精度を上げて行きたいと思っている。

８．謝辞

本研究は、ネットアドバンス社様に提供していただいた小学館様発行の「日本大百科全書」を対象に開発を行った。データの作成には野畑恵理子さん、竹内康介さんに手伝ってもらった。また、以下の方々とは議論を通じて刺激を受けた（順不同）。新山祐介さん、村上浩司さん（ニューヨーク大）高橋哲朗さん（奈良先端大）野畑周さん（シャープ）。ここに記して感謝の印とする。また、本システムはネットアドバンス社様が提供しているhttp://japanknowledge.comにおいて実験工房として動作している自然文質問応答システムでサービスされる予定である。

参考文献

QAC HP: http://www.nlp.cs.ritsumei.ac.jp/qac/index-j.html

TREC-QA HP: http://trec.nist.gov/data/qa.html

E. Agichtein and L. Gravano. 2000. “Snowball: Extracting relations from large plain-text collocations.” In Proc of the 5th ACM Intl. Conf. On Digital Library, pp85-94.

S. Brin. 1998. “Extracting patterns and relations from World Wide Web”, In Proc. of the WebDB Workshop-98, pp 172-183.

M. Fleischman, E Hovy and A. Ehcihabi: 2003. “Offline Strategies for Online Question Answering”, In Proc. of 41st ACL

D. Lin and P. Pantel: 2001. “DIRT – Discovery of Inference Rules from Text”, In Proc. of the 7th ACM SIGKDD, pp323-328.

D. Ravichandran and E. Hovy: 2002. “Learning Surface Text Patterns for a Question Answering System”, In Proc. of the ACL.

S. Sekine, K. Sudo, C. Nobata: 2002. “Extended Named Entity Hierarchy”, In Proc. of LREC 2002.

関根聡: 2003.「百科事典を対象とした質問応答システムの開発」言語処理学会第 9 回年次大会　発表論文集, pp637-640.

高橋哲朗、乾健太郎、松本裕治: 2004.「テキストから属性関係を抽出する」情報処理学会　自然言語研究会－164