E-023

インターネット上のヘッドラインニュースを情報源とした質問応答システムの構築

Development of a Question and Answer System using Headline News on the Internet as an Information Source

河野 安友未† 小林 一郎‡
Ayumi Kono Ichiro Kobayashi

1. はじめに

今日ではインターネットの普及により，多くの電子情報を収集することが容易となってきている．近年，最新の情報をある決まった形式で公開し，世界中で利用可能な資源として共有するという動きが広まっている．その情報公開の新しい形態の一つとして，インターネット上のサイトの見出しや要約などのメタデータを構造化して記述する RDF または XML 形式の一種である RSS（RDF Site Summary）がある [1]．RSS は，主にサイトの更新情報を公開するのに使われ，ニュースサイトなどは，RSS で記事のタイトルを配信するのに利用している．

本研究では，インターネット上の情報，特に時事ニュースをこの RSS から取得し，テキスト解析を行った結果をコンピュータに蓄積する．さらに，その蓄積された時事ニュースのイベントに対して，質問応答できるシステムの構築を行うことを目的とする．

2. システム概要

図 1 に構築したシステムの処理の流れを示す．

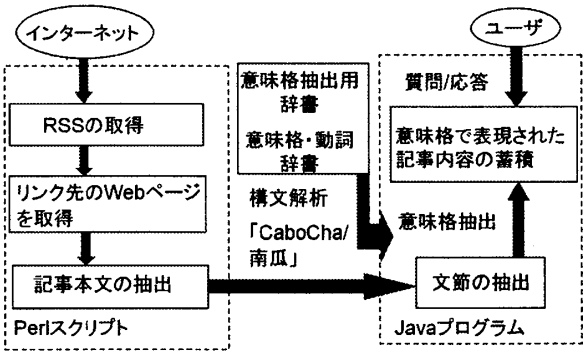

図 1: システムの処理の流れ

2.1 インターネット上のテキスト情報抽出

質問応答するための知識獲得のプロセスとして，インターネット上に公開されている RSS を読み取り§，その中に記載されている URL のリンク先のニュース記事の HTML ファイルを自動取得し，Perl スクリプトを用いてその中から記事本文のみを抽出してテキストファイルとして保存している．図 2 にテキスト抽出処理の流れを示す．

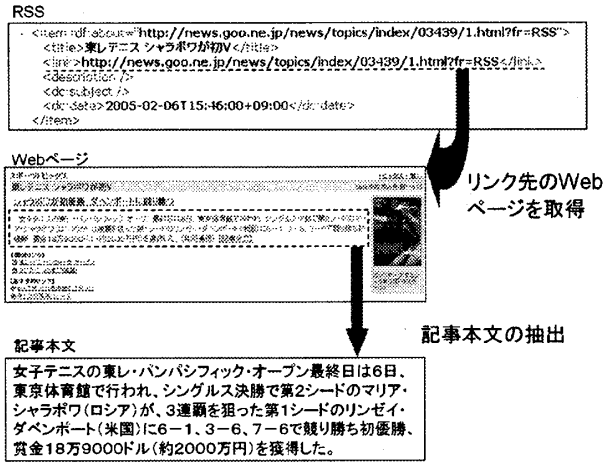

図 2: RSS からのテキスト抽出の流れ

2.2 テキスト解析

まず対象となるテキストを形態素解析，係り受け解析し，語の品詞情報，係り受けの情報を取り出す．このプロセスには奈良先端科学技術大学大学院で開発された構文解析日本語係り受け解析器「CaboCha/南瓜」[2] を使用した．解析した結果の中から，述語となる語の基本形を取り出し，その述語に係る文節を取り出す．文節が格を持つならば，格の種類を判別し，述語の基本形とそれがとる必須格，そしてその格に対する意味の制約（意味の選択制限）をまとめた意味格抽出用辞書を参照し，格がとる語の意味格（意味ラベル）を決定する．意味格（意味ラベル）とは，格がとる語の意味的制約のことを指し，本研究では，文献 [3] に基づき，9 種類の格「ガ格」「ヲ格」「二格」「カラ格」「ト格」「デ格」「ヘ格」「マデ格」「ヨリ格」に対して，35 の意味格を用意している．これに基づき述語および格による意味格の分類を行う．その一例を挙げると，たとえばデ格「○○で」において，述語と「で」の直前の語の持つ意味によって「で」の解釈が変わる．

教室で遊ぶ : 場所を示す
虫眼鏡で観察する : 道具を示す
風邪で休む : 理由を示す

格と意味格の関係例を表 1 に示す．

†お茶の水女子大学大学院人間文化研究科数理・情報科学専攻, Graduate Division of Mathematics and Computer Science, Graduate School of Humanitics and Sciences, Ochanomizu University
‡お茶の水女子大学理学部情報科学科, Dept. of Information Sciences Faculty of Sciences, Ochanomizu University
§今回，利用した RSS は http://news.goo.ne.jp/news/topics/で公開されているものを利用した．

表 1：格と意味格の対応関係表

格	用法	例文	意味格	意味ラベル
ガ格	動きや動作の主体を表す	会議**が**始まった	行為者	actor
ガ格	状態の対象を表す	兄はドイツ語**が**できる	行為の対象	target
デ格	出来事・動作の場所を示す	結婚式はホテル**で**行われた	場所	place
デ格	状態を表す用法	いいコンディション**で**臨んだ	状態	state
デ格	道具・手段を示す用法	赤鉛筆**で**下線を引く	手段（道具）	means-tool
デ格	原因を表す用法	太郎は風邪**で**学校を休んだ	原因理由	cause

このような取り得る述語を意味格と格ごとに分類してまとめたものが意味格抽出用辞書である．格と述語を組み合わせて記憶させておくことによって，選択制限がなされ意味格を適切に判別することが可能になる．具体的な意味格抽出用辞書の例を以下に示す．

意味格抽出用辞書「de-tool.dic」

```
描く
観察する
...
```

道具（tool）としての意味格をもつデ格（de）とその組み合わせのときに現れる述語の基本形が収められ，適切な意味格を判別する辞書となっている．この手法は一般的に「用例に基づく方法（example-based method）」[4] とみなされる．

意味格抽出用辞書の格および意味格ごとに収録された述語を利用し，文章中から格が抽出されれば，始めに文章中から取り出しておいた述語とマッチングをとる．述語が意味格抽出用辞書に収録されていれば，格の直前の語はその意味格と判別される．

実際に，構築したシステムによって例文「会議が始まった」という文を解析した結果を図 3 に示す．

```
EOS
* 0 1D 0/1 0.00000000
会議    カイギ  会議    名詞-サ変接続                  0
が      ガ      が      助詞-格助詞-一般               0
* 1 -1D 0/1 0.00000000
始まっ  ハジマッ 始まる  動詞-自立五段・ラ行    連用タ接続
た      タ      た      助動詞  特殊・タ 基本形  0
EOS
```

図 3: 「CaboCha/南瓜」による解析結果例

形態素解析，構文解析の結果，図 3 のような結果が得られるので，動詞「始まる」を取り出し，それに係る文節「会議が」に注目する．その文節の末尾に「ガ格」の「が　助詞－格助詞－一般」とあるので，動詞を引数として「ガ格」の辞書を参照する．ガ格の辞書には「ガ格」が取り得る意味格として actor と target の 2 種類が用意されており，「始まる」という語は actor の辞書に含まれているので，「ガ格」の前の語「会議」は actor と判別される．

これまで説明した手法を用いて，次の例文「太郎は双眼鏡で泳いでいる鳥を見た」を解析してみる．

Predicate　:　value=見た
actor　:　value=太郎
goal　:　value=双眼鏡で泳いでいる鳥

係り受け関係を基準にして意味格を抽出しているため，このように goal の値が「見た」に係る全体の文章として抽出されることになる．さらに「CaboCha/南瓜」では「双眼鏡で」の部分が「見た」に係るのではなく「泳いでいる」に係っていると判断されてしまう．

そこで goal として抽出された部分を，係り受けには注目せずに辞書を用いて意味格を抽出する．このときに使用する辞書は格関係と動詞ごとにまとめられた意味格・動詞辞書である．意味格・動詞辞書のデ格・動詞辞書を参照して「見る　双眼鏡」が含まれていれば「双眼鏡で見た」という係り受けと判断される．

表 2：意味格動詞辞書の例 dekakudoshi-tool.dic

動詞	意味格例
見る	双眼鏡
見る	虫眼鏡

表 3：意味格動詞辞書の例 dekakudoshi-place.dic

動詞	意味格例
見る	映画館
泳ぐ	池

もう一つの例文「太郎は池で泳いでいる鳥を見た」を解析してみる．

Predicate　:　value=見た
actor　:　value=太郎
goal　:　value=池で泳いでいる鳥

この例文でも goal の値が「見た」に係る全体の文章として抽出されてしまうので，意味格・動詞辞書を参照する．デ格・動詞辞書を参照して「見る　池」が含まれているかを確認する．「見る　池」は含まれていないので，含まれていなかった場合は goal の値の中に動詞がないかに注目する．すると動詞「泳いでいる」があるのでデ格・動詞辞書を参照して「泳ぐ　池」が含まれているかを確認する．含まれていれば「池で泳いでいる」という係り受けと判断される．

この意味格・動詞辞書は意味格抽出用辞書では抽出しきれない文などにも対応できる.「映画館で見る」「双眼鏡で見る」など「見る」に対してデ格がとりえる意味格は「道具」と「場所」があるため，意味格抽出用辞書のみでは意味格を一意に定めることができないが意味格・動詞辞書を用いれば意味格を正確に抽出することができる．しかし，意味格抽出用辞書を用いても一意に意味格を定めることのできる文章は数多く存在する．システムでは意味格抽出用辞書で定め切れなかった意味格を，意味格・動詞辞書を用いて正確に抽出するという辞書の二段組構造になっている．二段構造になっているのにはもう一つ理由があり，意味格・動詞辞書に述語と意味格例を人手で記述していくのにはコストがかかりすぎる．意味格抽出用辞書を用いてある程度意味格の具体例を抽出し，意味格・動詞辞書に追加していくことによって大幅なコスト削減を実現する．

テキスト解析により抽出された意味格は，意味格が付与される XML 形式でファイルに保存される（図 4）.

```
- <sentence>
    <Predicate value="始まった" />
    <actor value="会議" />
  </sentence>
- <sentence>
    <Predicate value="できる" />
    <actor value="兄" />
    <target value="ドイツ語" />
  </sentence>
```

図 4: 抽出された意味格（XML 形式）

2.3 抽出された内容に対する質問応答

質問文も同様に「CaboCha/南瓜」によって構文解析される．それによって得られた文節の情報から述部を取り出し，XML ファイルとして蓄積された記事内容の述部のとる値（Predicate の属性 value が取る値）とマッチングを取る．マッチングが取れた場合，現在のシステムでは，ユーザからの質問文中にキーワードとなる語を見つけ出し，それに対する意味格がとる値を用意されている回答文のテンプレートのスロットに埋め込み，ユーザに回答するようになっている（図 5 参照）.

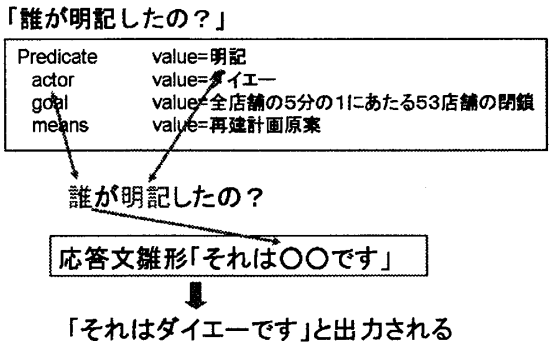

図 5: 質問に対するシステムの回答文生成

実際にユーザの質問に対してシステムが回答している例を図 6 に示す.

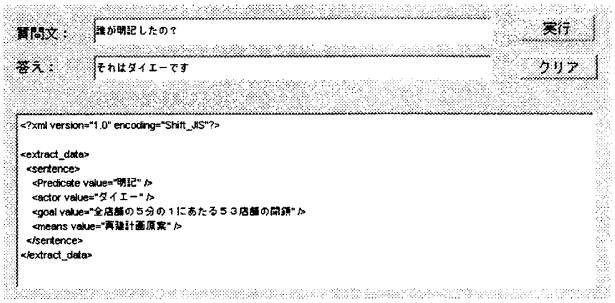

図 6: 質問に対するシステムの応答例

3. まとめと今後の課題

本研究では，インターネット上に公開されている情報を知識源とし，質問応答できるシステムの開発を行っている．システム開発の現状としてインターネット上からテキスト情報を自動取得し，意味格を抽出，意味格に対しての質問応答を行うという部分は完成した．現段階で 35 種類すべての意味格が抽出可能である．今後は意味格抽出辞書を充実させ，より多くの文章から意味格を抽出できるようにするとともに，抽出した意味格を意味格・動詞辞書に自動的に追加していき抽出精度の高いシステムの開発を進める．さらに抽出された意味格・動詞辞書の意味格の部分にシソーラスを用いて様々な表現の文章からも柔軟に意味格を抽出できるようにする．現在は抽出した意味格をそのまま DOM によって管理される XML 形式のファイルとして出力しているが，SemanticWeb の枠組みにおいて知識表現として用いられているメタデータ記述言語 RDF や RDF スキーマの形式で出力するようにする．これにより，推論機能を取り入れ時系列でみたときのヘッドラインニュースの情報「最近の日経平均株価の動向は？」などの質問にも柔軟に回答できるようにするつもりである．

参考文献

[1] W3CDTF, Misha Wolf and Charles Wicksteed, Date and Time Formats , 1997-12-15, W3C Note , http://www.w3.org/TR/NOTE-datetime

[2] 奈良先端科学技術大学大学院松本研究室, 日本語構文解析器「CaboCha/南瓜」, http://chasen.org/ taku/software/cabocha/

[3] 益岡隆志, 田窪行則, “基礎日本語文法”, くろしお出版,1992.

[4] 長尾 真編, 岩波講座ソフトウェア科学 15 “自然言語処理”, 岩波書店, 1996.

[5] THE WEB KANZAKI, メタ情報とセマンティック・ウェブ,2001-09-16, http://www.kanzaki.com/docs/sw/

[6] 斎藤信男, 荻野達也, “セマンティック Web 入門”, オーム社,2004.