1ZA-02　音声対話システムのためのWebからの知識収集

藤澤正樹　児玉 浩之　荒木 雅弘　新美 康永
京都工芸繊維大学工芸学部電子情報工学科

Abstract

対話システムでは，人間ーコンピュータ間で種々の問題解決を行うためにシステムが知識情報を持ち合わせている必要がある．こういった知識の情報源として本研究では，インターネット上のWebページに着目する．知識情報データベースとして，HTMLで記述されたWebページを汎用データ記述言語として注目されているXML形式に変換することを試みる．変換にはWebページの構造情報と，キーワードのシソーラス上での関係を利用する．実際に，提案した方法にしたがって情報を収集し，その収集効率について検討した．

1．はじめに

対話システムでは，さまざまな問題の解決を行うために知識情報を持ち合わせている必要がある．一般にこれらは個別のシステムにあわせて設計・準備されている．しかし，準備された知識情報は時間の経過と共に内容が古くなっていく．このため定期的な知識情報の更新が必要となってくる．知識情報の準備や更新は，さまざまな情報源からデータの収集を行い，個別に必要な形式へと変換してゆかなければならない．

本研究では，これらの作業を軽減するために，対話システムで利用可能な知識情報の自動生成を試みる（図1）．

2．知識情報の収集

Web上には多くの情報が存在し，それは常に更新されつづけている．本研究では，こうしたWebの特徴に注目し，その中で主流をなすHTMLを知識情報収集のターゲットとしている．これまでにも，HTMLから情報を収集する試みが行われてきた[1]．本研究では情報の収集にとどまらず，それらを知識情報として構造化することを試みている．しかし，HTMLは本来，構造をあらわすべきタグが表示のみを目的として誤用されている場合が多く，構造的なデータとはみなされない．そのため，そこから知識情報を収集するためには内容や前後関係などを参照するほかに手段がない．

一方，XMLは文書構造定義が厳密であり，タグ名が内容を表すことから，音声対話システムの知識情報として利用することができる(図2)．そこで，HTMLで表現されたWebページを対話システムの知識情報として利用可能なXMLに変換するアルゴリズムを提案する．

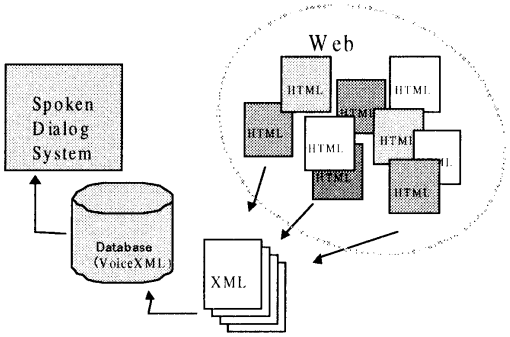

図1．対話システムと知識表現

2.1　情報ブロックの決定

HTMLに記述されている情報を収集するためには，どこに何の情報が記述されているかを知る必要がある．しかし，HTMLタグは，本来構造を示すものが表示目的に転用されたり，閉じタグが省略可能であることからHTMLタグのみでは内容を構造化することは難しい．本研究では知識情報収集の第一歩として，HTMLからタグの階層情報を抽出し，用意したキーワード群を手がかりに，特定の情報が記述されていると予測される部分を情報ブロックとして分割．抽出する．

まず，Webページ上での一般的なHTMLタグの使用法から，知識収集に有用な階層情報を構成しないタグを取り除き，その上で残ったタグに囲まれた部分ごとに階層付けを行う．たとえば

```
客室情報
<DL>
    <DD> シングル
    <DD> ダブル
</DL>
```

のような場合，抽出した階層構造は

Knowledge acquisition from the Web for Spoken Dialog Systems
Masaki Fujisawa
Department of Electronics, Kyoto Institute of Technology

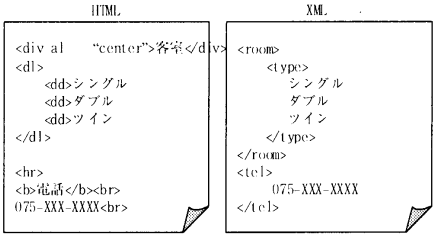

図 2.HTML と XML

```
<LAYER RANK=" 0" > 客室情報
    <LAYER RANK=" 1" >
        <LAYER RANK=" 2" > シングル </LAYER>
        <LAYER RANK=" 2" > ダブル </LAYER>
    </LAYER>
</LAYER>
```

となる. このように分類した階層の一塊を階層ブロックと呼ぶ. 前述の例でいうならば, 一組の <LAYER> タグで囲まれた部分が一つの階層ブロックをなしていると考えられる.

次に収集したい情報の種別をあらわすキーワード (情報項目キーワード) と, 実際に記述される具体的な内容をあらわすキーワード (情報内容キーワード)をそれぞれ用意する. これらは, あらかじめ定義したDTD(Document Type Definition)にあわせて設定する. キーワード群の生成は, シソーラス[2]を用いて類義語などをリストアップしているが, 現在のところその多くは手作業で行っている.

用意した情報項目キーワードで, 階層ブロックごとにマッチングを行う. 次にマッチした階層ブロックやそれに隣接して続くブロック, 一つ上の階層の親ブロックなどについて, 順次情報内容キーワードをマッチングして行く. その結果, キーワードがマッチした数によって, 1つまたはいくつかの階層ブロックを一つの情報ブロックとして決定することができる.

こうして用意された情報ブロックは, 特定の情報が含まれる部分として, 次の処理に用いられる.

2.2 知識情報の抽出

知識情報データベースとしてまとめるためには, 前節の処理で決定された情報ブロックから DTD の定義にしたがって, XML へと情報を埋め込んでいく必要がある. ここでは, キーワード群を決定する際に定義した情報の種別ごとに, その内容が文で表されるのか, 価格のように数字と記号の列で表されるのかなどを, あわせて定義しておくことにより, 情報ブロック内の形態素解析結果とあわせて知識情報の収集を行っている.

ホテルの部屋種別などのように, 収集する情報が単語と呼ばれるような単位であらわされるような場合, 情報内容キーワードにマッチした単語をピックアップすることで, 決まりきった情報を抽出することができる. また, 文章で表されるような場合, 情報内容キーワードが多くマッチする一文を抽出することにより, 内容を取り出すことができる.

表 1. 知識収集の効率

ホテル	名前	住所	電話	FAX	交通 (自動車)	交通 (鉄道)	部屋種別	料金	チェックイン・アウト
ホテルA	×	×	△	△	—	○	○	○	○
ホテルB	×	×	△	×	△	×	○	△	△
ホテルC	×	×	×	—	△	△	○	×	×
ホテルD	×	×	△	△	—	—	○	△	△

○…抽出成功, △…余分な情報を含んでいるもの, ×…抽出失敗, －…サイトに掲載されていない情報.

3. 実行結果

2 章で述べた方法にしたがって, 実際のHTMLから知識情報の収集を試みた. 評価に用いたWebサイトとしては, 京都府のホテルサイト 4ヶ所分を使用した.

結果を表1に示す. 余分な情報は含むものの、必要な情報の抽出が成功したのは, 全体の63.8%であった.このうち、余分な情報を含むことなく完全な抽出が行われたのは ,30.4%(全体の 19.4%)であった .

4. おわりに

今後は, 抽出が不完全であったものを改善するために単語と呼ばれるような情報か文章だけではなく, 新たな表現のタイプを定義して精度向上を目指す. また, マッチングによる情報抽出も, 概念的に近いものなどをうまく収集できるような仕組みを検討したい.

参考文献

[1]Stephen Soderland “Learning to Extract Text-based Infomation from the World Wid Web” InProceedings of Third International Conference on Knowledge Discovery and Data Mining(KDD-97)

[2]池原 悟・宮崎正弘・白井 諭・横尾昭男

中岩浩巳・小倉健太郎・大山芳史・林 良彦 “ 日本語語彙体系” 取扱説明書 (1999)