Ｗｅｂを情報源としたＱ＆Ａシステムの検討

山田一郎[*1]　柴田正啓[*1]　金淵培[*1]　崔杞鮮[*2]
[*1] ＮＨＫ放送技術研究所, [*2] Korea Advanced Institute of Science and Technology
E-mail: {yamada.i-hy, shibata.m-mg, kimu.y-go}@nhk.or.jp, kschoi@cs.kaist.ac.kr

1 はじめに

近年、放送局では多種多様なコンテンツを扱うようになった。NHKにおいても、放送された映像・音声コンテンツや、番組台本、ニュース原稿などのテキストコンテンツを電子化し、容易に利用できる形で大量に蓄積するシステムが整備されつつある。そこで、我々は、これらのコンテンツを学校教育に活用するマルチメディア教育支援システムの研究を進めている[1][2]。

このシステムでは、ネットワーク上の仮想教室で、遠隔地にいる複数の生徒が話し合ったり、NHKの映像データやテキストデータにアクセスしたりしながらグループ学習を行う。仮想教室には、生徒のグループ学習を支援するエージェントが参加する。エージェントが持つ情報として、用語とその説明が記述された用語辞書が重要な役割を果たす。我々はこれまで、用語辞書を自動生成するために、NHKの放送用読み原稿として利用されるニュース原稿を解析して重要な用語とその説明文を抽出する研究を進めてきた[3]。しかし、ニュース原稿のみでは、情報量に限りがある。そこで、大量の情報が蓄積・更新されているインターネットを情報源とし、エージェントが利用できる辞書情報を自動抽出する手法を提案する。

従来、藤井らは、文書表現とHTMLレイアウトに基づいてWebから用語説明を抽出する手法を提案している[4]。この手法では、約20万語の用語に対する説明が自動収集され、複数の用語説明を分野や語義に基づいて分類することも行われている。評価では、既存の事典よりも網羅性が高く、質も実用レベルという興味深い結果が得られている。また桜井らは、Webから抽出した用語説明を、文書表現に基づいて、その役割別に分類している[5]。これらの従来手法における文書表現は、「XとはYです」「XをYと定め」といった、主部—述部もしくは、動詞の格といった関係を利用している。我々の提案する手法では、それらの表現よりWebコンテンツにおける用語説明での頻度が高い連体節（用語を修飾する従属節）に注目して処理を行う。

本稿では、検索エンジンのGoogle[10]を利用し、対象用語をキーワードとして検索された上位10個のWebページから、対象用語と、その用語を修飾する節を抽出し、その連体節の用語に対する役割を推定する手法を提案する。ここでは役割として、語の意味記述を体系付けたQualia Structure[6]で定義された10種類の関係を用い、最大エントロピー法による学習を利用して、連体節の役割を分類する。最後に、生徒からの質問に対し、Webを検索し質問の答えとして最適な文をリアルタイムで答えるマルチメディア教育支援システムのQ&Aシステムを紹介する。

2 Webにおける用語説明

Webページで用語を説明するパターンには、次の4種類が考えられる。

A) 定義型リスト形式、もしくは、ボールドタグなどにより用語が見出し化
例：
```
<DL>
  <DT>情報家電
  <DD>次世代の高速インターネットに対応した家電製品
</DL>
```

B) 用語にリンク先が付加
例：
```
<A HREF="#kaden">情報家電</A>
<A NAME=kaden>次世代の高速インターネットに対応した家電製品</A>
```

C) 文で説明
例：情報家電は、次世代の高速インターネットに対応した家電製品です。

D) 連体節で説明
例：次世代の高速インターネットに対応した情報家電は、～

これらの例では、全て、用語「情報家電」の説明を行っている。Googleの検索結果に対して、各パターンがどれくらいの頻度で出現しているか、手作業により調査を行った。対象用語として、2002年11月のニュース原稿に出現した最新の時事用語71語を利用した。検索結果上位10ページに1回以上、上記パターンにより用語を特徴付ける十分な説明が出現した割合を表1に示す。

この結果では、47.9%の用語に対してパターンDによる説明が存在しており、これらの用語説明パターン中では最も多く出現していた。全てのパターンを合わせる事により、検索上位10ページから60%以上の用

表1．Webにおける用語説明の出現割合

パターン	説明の存在割合
A	21 / 71 (29.6 %)
B	10 / 71 (14.1 %)
C	16 / 71 (22.5 %)
D	34 / 71 (47.9 %)
全体	43 / 71 (60.6 %)

語に対しての説明が抽出可能であることがわかった。

パターンA～Cについては、文献[4][5]により考察が行われているので深くは言及しない。本稿では、最も出現頻度が高い連体節の解析手法を提案する。

3 用語説明の抽出

本手法では、最初に、対象用語を検索キーとしてGoogleの検索を行い、その結果のWebページから、用語を含む文を抽出する。次に、抽出された文に、前節の用語説明パターンが存在するかを判定する。抽出された連体節を利用して説明文を生成し、最後に、説明文の役割を判定する。以下に各処理について述べる。

3．1 連体節の解析

HTML文書では、作成者は表示レイアウトを意識して、文の途中でも改行タグを挿入することがある。そのため、文の区切りの抽出が問題となる。本手法では、</P>等のHTMLタグと、句点を手掛かりとして文の区切りを判定し、不要なHTMLタグを除去して、文を再構成する。この際、付加されていたHTMLタグを基に用語の説明パターンA、Bの有無を判定する。

次に用語が含まれる文のみを構文解析し、用語の説明パターンC，Dの有無を判定する。文中に、「XはYです」というパターンがある場合、Yの部分を用語Xの説明として抽出する。用語に係る連体節がある場合は、連体節の部分を用語説明として抽出する。この処理では、十分な情報量を持つ説明を抽出するために、動詞が含まれている連体節のみを処理対象とする。

ここで抽出された連体節は、動詞の連体形で終わっているため、文として不完全である。そこで、用語の上位概念語を抽出して、この連体節と統合する。前節の例でも、パターンDの例では「次世代の高速インターネットに対応した」の部分が用語に係る連体節となっている。用語「情報家電」の上位概念として「家電製品」が抽出できれば、「次世代の高速インターネットに対応した家電製品」という他のパターンと同様の説明文が抽出できる。上位概念抽出処理は、以前我々が考案した抽出手法を利用した[3]。この手法では、「と呼ばれる」などの定型的な言い換え表現や、パラフレーズ構造、形態素情報などを利用して上位概念を抽出している。抽出処理の結果、75.7%の用語に上位概念が付加でき、適合率が95.7%と良好な結果が得られている。上位概念語が得られなかった残りの用語は、「もの」「こと」という一般的な単語をその上位概念とする。

3．2 説明文の役割判定

抽出された説明文の用語に対する役割は、様々な種類が考えられる。本手法では、この役割を、語の意味記述を体系付けたQualia structure[6]を参考にして分類する。Qualia structureは、欧州における１２の言語で相互参照できる語彙の意味体系を構築することを目的としたプロジェクトであるSIMPLE (Semantic Information for Multifunctional Plurilingual LExicons[7])で採用された語彙体系である。Qualia structureでは語の指示対象（概念）の意味記述の要素として、以下の４つの役割を定義している。

A)Formal role：語の指示対象と他を区別する情報（≈Is_a関係）
B)Constitutive role：指示対象が持つ内部的な性質・構成要素
C)Telic role：指示対象が持つ典型的な機能・目的
D)Agentive role：指示対象の起源、指示対象が引き起こす事象

この意味記述の役割を、用語の説明部分の役割に当てはめる。下記の例では、下線部の説明が、鍵括弧で囲まれた用語におけるQualia structure(A～D)のそれぞれの役割となる。

A)ビル街や舗装された道路に囲まれた都会の気温が上昇する「ヒートアイランド現象」が、～
B)この他、細長い花弁を水牛の角のように張る「スカホセパラム」の仲間など～
C)森林の整備や緑化の推進を目的とする「緑の募金」を、～
D)ハタミ大統領が提唱する「文明間の対話」～

5W1Hの関係や特徴となるConstitutive roleについては、さらに細分化された下記の役割を利用して分類する。

E) Time　　：時間
F) Location　　：場所
G) Instrument　　：道具
H) Contain　　：包含
I) Is_a_member_of　：メンバー
J) Is_a_part_of　：部分
K) Constitutive　：E)～J)以外のconstitutive role

用語の説明部分の役割が、選択した10個のQualia Structureのいずれに該当するかを推定するために、説

明部分と用語の表層的な特徴を手掛かりとする、最大エントロピー法による学習法を利用する。学習における素性は、連体節中の用語を直接修飾する動詞の標準表記・時制・態・完了形の有無、動詞の格、その格が含まれる節の自立語表記、自立語の属性、用語の属性とする。ここで自立語と用語の属性は、IREX 固有表現抽出タスク[8]により付加された８つの固有表現(組織名、人名、地名、固有物名、日付表現、時間表現、金額表現、割合表現)とし、抽出アルゴリズムは内元らの手法[9]により、あらかじめ判定した結果を利用する。図１に素性の例を示す。

	千葉県柏市を	ホームタウンと	する	「柏レイソル」
表記	千葉県柏市	ホームタウン	する	
属性	地名	その他		組織名
格助	を格	と格		
動詞の時制・態・完了形の有無	現在形・能動態・無			

図 1. 付加する素性例

この例では、“千葉県柏市をホームタウンとする「柏レイソル」～”という文から取り出された連体節「千葉県柏市をホームタウンとする」と用語「柏レイソル」の関係を推定するために、表記（千葉県柏市、ホームタウン、する)、属性（地名、その他、組織名)、助詞（を格、と格)、動詞の特徴（現在形、能動態、完了形無し）といった特徴を利用している。

この素性により、連体節の役割が Qualia Structure のいずれに該当するかを推定する。一つの用語に対して複数の同じ関係を持つ連体節が抽出される場合は、どの連体節が、その役割として相応しいか優先順位をつける必要が生じる。そこで、抽出された全ての説明集合に類似する特徴を持つ連体節ほど、ノイズが少なく重要な情報と判断する。この処理では、含まれる単語をベクトルの項として、全ての説明を特徴付ける重心ベクトルと、各連体節の特徴ベクトルを定義し、その内積値の降順に順位付けを行う。最後に上位概念を付加して役割ごとの説明文を生成する。

4 実験

前章までに提案した手法の有効性を検証するために、ニュース原稿に出現した時事用語を対象に、役割別の説明文を Web から抽出する実験を行った。学習データは、ニュース原稿から 5,481 組の連体節と用語の組を無作為に抜き出し、それぞれに Qualia structure の役割分類を手作業で与えて作成した。この際、学習データに少数しか出現しない素性はノイズとなる可能性があるため、出現頻度が 8 回以上観測された素性のみを用いている。推定処理の結果、各役割に所属する確率値が出力される。処理結果の一部を表２に示す。2002 年 11 月に出現した 376 個の時事用語から抽出された 357 個の説明文に対する役割分類の評価結果を表３に示す。この役割分類の結果は、ニュース原稿を対象として行った実験結果[3]（適合率 81:4%、再現率 76.0%）と比べて悪い。学習データが、使用される単語や表現が限られたニュース原稿であり、Web データは様々な表現が使用される均整のとれていないテキストである事が原因と考えられる。しかし、判定閾値を変更することにより、適合率は向上可能であり、エージェントが持つ辞書情報として利用できる。

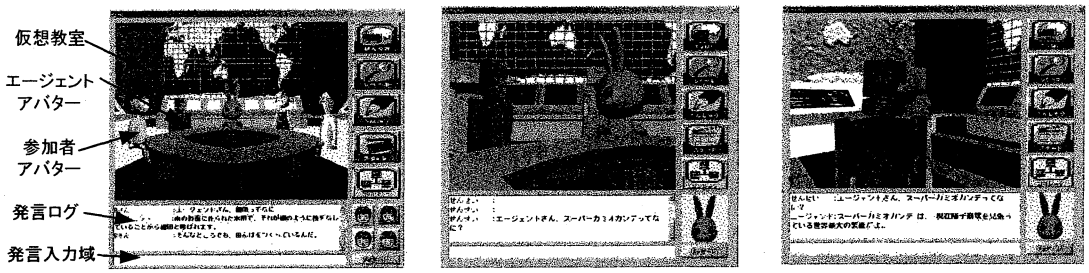
表 3. 用語説明文の役割判定結果

適合率	69.4%
再現率	54.6%

formal の役割として正確に分類された説明が獲得できた用語数は 52 個（13.8%)、telic は 15 個（4.0%)、agentive は 58 個（15.4.%)、E～K までのすべての constitutive は 28 個（7.4%）であった。125 個(33.2%)の用語に１個以上の説明文が抽出された。

表 2. Web から抽出された用語、用語説明文、役割

用語	説明	役割	確率値
スーパーカミオカンデ	巨大なタンクでニュートリノをとらえるもの	formal	0.501
スーパーカミオカンデ	東京大学宇宙線研究所の神岡宇宙素粒子研究施設(岐阜県神岡町)にある巨大観測装置	location	0.892
スーパーカミオカンデ	作業中に光センサー数千本が破損するという大事故にあったニュートリノ観測施設	agentive	0.525
テーラーメイド医療	遺伝情報を基にした個人個人にあった予防・治療を可能とする医療	formal	0.576
テーラーメイド医療	我々が提唱する医療	agentive	0.668
全国おさかなまつり	食料供給など漁業のもつ役割や環境問題の啓蒙も目指すまつり	telic	0.988
津波ハザードマップ	平成14年11月22日に設置された地図	time	0.943



複数の生徒アバターとエージェントが参加するネットワーク上の教室　　アバター(うさぎ)がエージェント(ロボット)に質問　　エージェントがWebから情報を探し出し返答

図2. マルチメディア教育支援システムのユーザーインターフェイス

5 教育支援システムにおけるQ&A

Webから抽出される用語の説明文とその役割を、マルチメディア教育支援システムにおけるエージェントの辞書情報として利用する。このシステムでは図2に示すユーザインターフェースを用い、参加する生徒がチャットを行うように文字をコンピュータに打ち込むことにより、生徒の化身となるCGアバターを通して話し合う。

エージェントは、生徒の会話内容をモニターし、エージェントに対する呼びかけをキーとした質問文章に対して、表層的なテンプレートマッチによる解析によりキーワードを抽出して必要な情報の検索を行い応答する。生徒から、「エージェントさん、"用語"って何?」という質問があった場合、エージェントはWebを検索し、用語のFormalの関係にある説明文を返答する。他にも、「どういった特徴?(Constitutive)」、「何のため?(Telic)」「どこ?(Location)」などの質問に対しても、対応する関係にある説明文を利用して返答できる。連体節は、用語の説明が簡潔に記述されているため、それを利用したQ&Aシステムでも短い文で返答でき、このようなシステムに適した情報と考えられる。この際、返答までに要する時間が問題となる。実験で利用した11月の時事用語に対する処理に要した時間を調査した結果、説明を抽出できた場合は平均13.9秒であった。所要時間の内訳は、Webページの獲得に平均5.3秒、構文解析に7.7秒、役割判定に0.9秒であった。所要時間は実験環境にも依存するが、リアルタイムでQ&A処理を行うためには、今回実験した10ページ程度の処理が限界と考えられる。本システムでは、全ての処理に30秒以上要する場合は、その時点での処理結果から返答している。

6 まとめ

本稿では、Webを対象とした用語の説明とその役割を抽出する手法を提案した。さらに、生徒とエージェントの対話を行う仮想教室システムへの応用を図った。Webページの情報は、使われる単語や表現にバラつきが多いテキストであるため、依然、解析における問題点が多いが、一つの情報源となりうることを確認した。

今後、今回の実験で発生したWeb解析上の問題点を解決するとともに、Webから人物の説明を抽出する手法などへ発展させていく予定である。

【参考文献】

[1] 住吉ほか「学習コミュニティの対話を支援する仮想教室のシステム化」FIT2002 情報科学技術フォーラム, V-41(2002)

[2] 住吉ほか「新しい教育放送サービスのための映像検索システム」映メ学会誌 Vol. 57, No.2, pp253-261 (2003)

[3] 山田ほか「ニュース記事に出現する用語と説明文の意味関係自動獲得」情処学会研究報告, NL152-21, pp145-152 (2002)

[4] 藤井ほか「Would Wide Webを用いた事典知識情報の抽出と組織化」信学会論文誌,Vol. J85-D-II No.2, pp300-307 (2002)

[5] 桜井ほか「ワールドワイドウェブを利用した用語検索の実現」情処学会研究報告, NL137-4, pp23-29 (2000)

[6] Pustejovsky, J.: "*The generative Lexicon.* Cambridge", MA: MIT Press. (1995)

[7] http://www.ub.es/gilcub/SIMPLE/simple.html

[8] http://www.cs.nyu.edu/cs/projects/proteus/irex/

[9] 内元ほか:"最大エントロピーモデルと書き換え規則に基づく固有表現抽出", 自然言語処理 Vol. 7, No. 2, pp63-90 (2000)

[10] http://www.google.co.jp