新規情報収集エージェントによる進化型ウェブ大百科事典の構築

Building the Evolvable Web Encyclopedia by Information-gathering Agents

藤井　敦 [1)]　伊藤　克亘 [2)]
Atsushi FUJII　Katunobu ITOU

1) 図書館情報大学（〒305-8550 つくば市春日 1-2　E-mail: fujii@ulis.ac.jp）

2) 産業技術総合研究所 情報処理研究部門（〒305-8568 つくば市梅園 1-1-1　E-mail: itou@ni.aist.go.jp）

ABSTRACT. The World Wide Web, which contains a surprisingly much volume of up-to-date information, is a promising source to obtain encyclopedic knowledge for new and technical terms. In this project, we developed a system (software agents) to produce encyclopedias from the Web. Our system first searches the Web for pages containing a term in question, and analyzes HTML layout of pages to extract paragraph-style fragments that describe the term. Finally, only high-quality fragments are selected and organized based on domains in a resultant encyclopedia. We used our system to build an encyclopedia including approximately 200,000 entries. We also performed a number of experiments, in which human assessors and a large number of anonymous subjects evaluated the quality of our encyclopedia. The results showed that our project potentially rivals existing hand-crafted encyclopedias and Web search services.

1　背景

インターネットを利用して誰もが手軽に情報を発信できるようになったことを主な要因として，情報洪水と呼ばれるほど大量の情報が氾濫するようになった．このように日々増え続ける情報に囲まれた生活環境の中で，我々は未知の言葉に日常的に遭遇する．

知らない言葉や事柄について調べるための情報源として，昔から国語辞典や百科事典がある．しかし，既存の辞典や事典は頻繁に改定されるわけではないため，日々生み出される新しい事柄や専門技術に関する言葉は収録されていないことが多い．また，既存の言葉に対する新しい定義は収録されておらず，既存の定義ですら全て収録されているわけではない．そこで，冊子体・電子版といった媒体の形式に拘らず「量的問題」が発生する．

それに対して，World Wide Web には専門性が高い最新情報が存在するため，Web 上の検索エンジンを使って調べものをすることは日常的になってきている．しかし，既存の検索エンジンを使うと検索結果には膨大な数の不要なページが含まれるため，ある用語に関する事典情報だけを選択的に取得することは困難である．また Web には統制がないため，誤字，誤解，嘘などの低品質の情報も存在する．すなわち「質的問題」が発生する．

筆者らは Web を事典的に利用することを目的として，量質ともに優れた事典を構築するためのシステムを提案し，改良を重ねてきた [3, 4, 5, 13, 14] ．本システムは Web ページに含まれる良質な用語説明を選択的に抽出し，専門分野に基づいて分類することで事典を自動構築する．その結果，ユーザは検索語の説明を分野ごとに閲覧することが可能となった．

本システムの性能を評価するために，100 語程度の情報処理用語を用いて主観評価を行った．さらに，情報処理技術者試験の用語問題を解く応用タスクによって客観的に評価した．どちらの評価でも概ね良好な結果を得ている．しかし，当実験に用いた対象用語数は比較的限定されていたため「既存の事典を凌駕するような大規模で多種多様な用語を含む事典を Web から自動構築することは可能か?」という問いに対して断定的な答えを出すことは困難であった．

2　目的

本事業では，Web から新語に関する説明情報を継続的に収集し，組織化することで，常に進化し続ける大百科事典の実現を目的とした．そのために，筆者らが研究開発した事典構築のプロトタイプシステムに基づいて，基盤となるソフトウェアエージェント群を開発した．このような開発は，それ自身，単体でも価値ある事業である．他方において，当該エージェントを運用するためには計算機やネットワークに関する法外なコストを必要とするため，エージェントプログラムを一般に供しても，誰でも簡単に実用規模の事典を構築できるとは限らない．

そこで，本事業ではエージェントの開発だけでなく，それを実際に運用して事典（コーパス）[*1]を構築した．既存の事典は数千から数万語程度の収録語数なので，本事業では百万語という前人未踏の事典を目指し，平成 13 年度は約 20 万語の事典を構築して評価を行った．

以下，3 章で事典の構築及び検索システム（事典システム）について説明し，4 章では実際に構築したコーパスの分析を通してシステムの性能を評価する．5 章では不特定多数の被験者を対象に行ったモニタ調査の結果について報告し，6 章で本事業の成果について総括する．

3　事典システム

(1)　概要

筆者らが開発したシステムは，与えられた用語集に対する事典を構築するエージェント群と，構築されたコーパスをユーザインタフェースによって検索するサービス機能を含んでいる．以下，これらをまとめて「事典システム」と呼ぶ（図 1）．

[*1] コーパスとは言語データを体系的に集めたものである．本稿では，コーパスを実際に構築した事典と同義で用いる．

(2) 事典コーパスの構築

図 1 に基づいて事典コーパスの構築手法について説明する．対象用語集が与えられると，個々の用語に対して「検索」「抽出」「組織化」という 3 つの処理が順番に実行される．本事業では，それぞれの処理を実行するソフトウェアエージェントを開発した．

検索処理では，既存の Web 検索エンジンと同じように，対象用語を含むページを網羅的に取得する．原理的には，既存の検索エンジンを利用することが可能である．しかし，多数の用語を用いた大規模な実験を行うためには，通信などのコストが膨大になるため，独自にページ収集を行い，検索エンジンを実装した．主に日本のサイトから約 3 千万ページを収集した．これらのページを「茶筌」[*2]を用いて形態素解析して単語単位で索引付けを行い，キーワード検索を可能とした．

しかし，用語説明はページ全体ではなく一部であることが多いので，抽出規則によって特定の領域を抽出する．ここで，HTML のタグ構造を利用してページのレイアウトを解析し，対象用語を含む領域（段落）を抽出する．さらに，用語解説のページ特有のレイアウト（見出し語の強調やリンクの使用）を持つ段落も抽出する．

この段階では「対象用語が説明されている可能性が比較的高い段落」が収集されているだけである．そこで，組織化処理によって，対象用語が適切に説明されている度合が高い段落を優先的に取得し，さらに既存の事典と同じように語義や分野に応じて分類を行う．実際には，語義と分野は関連していることが多いので [11]，分野を分類することで間接的に語義を区別する．

ここで，対象用語が関連する一つ以上の分野があらかじめ分かっていると仮定する．組織化処理の目的は，関連分野 c に対して最適な用語説明 d を選択することである．確率的手法の観点から捉えれば，c に対して $P(d|c)$ を最大化する d を選択することに相当する．

実際の処理では，まず全ての分野に対して $P(d|c)$ を計算し，$P(d|c)$ の値がある閾値以上の d だけを選択する．その結果，対象用語が関連する分野と適切な用語説明を同時に特定することができる．そこで，はじめに仮定したように，対象用語が関連する分野をあらかじめ知っておく必要はない．

ここで，ベイズの定理によって式 (1) が成り立つ．

$$P(d|c) = \frac{P(c|d) \cdot P(d)}{P(c)} \tag{1}$$

式 (1) の右辺において，分母 $P(c)$ は定数として扱い，分子のみが組織化処理の中核である．$P(c|d)$ は用語説明 d が分野 c に関連する度合を定量化し，$P(d)$ は d が用語説明（事典情報）として妥当である度合を定量化する．両者をそれぞれ「分野モデル」「事典モデル」と呼ぶ．

分野モデルを実装するために（株）ノヴァ[*3]の機械翻訳用「専門語辞書」（19 の専門分野に対して合計約 100 万の日英対訳を収録）から抽出した日英項目を利用し，既存の文書分類手法 [7] によって $P(c|d)$ を計算した．

事典モデルを実装するために，用語説明としての妥当性について検討した．まず，言語的な妥当性がある．対象の用語について説明していない抽出結果は排除する必要がある．次に，情報の信頼性に関する妥当性がある．一般の出版物に比べると，Web ページは誤りや虚偽を含むことが多い．そこで，言語的に妥当であっても，信頼性が低い用語説明は排除しなければならない（尤もらしい嘘をつくことは可能であるため）．さらに，抽出処理で利用した HTML のレイアウト構造に基づいて，用語説明らしい段落を特定することもある程度可能である．

以上より，式 (2) に示すように，事典モデルを言語モデル $P_L(d)$，信頼性モデル $P_R(d)$，レイアウト（構造）モデル $P_S(d)$ に分解する．

$$P(d) = P_L(d) \cdot P_R(d) \cdot P_S(d) \tag{2}$$

言語モデルを作成するために「茶筌」を用いて日立デジタル平凡社「CD-ROM 世界大百科事典」（約 8 万語収録）を単語に分割し，CMU-Cambridge ツールキット [2] を用いて単語トライグラムを学習した．ここで，説明対象用語の表層的な違いは重要ではないので，見出し語を事前に共通の特殊記号に置換した．また，d に含まれる対象用語も同様の特殊記号に置換する．そこで，事典によく現れる表現（「X とは」や「と呼ばれる」など）を含む段落ほど高いスコアが与えられる．

さらに，信頼性モデルを実装するために，Google [*4]でページの順位付けに使用されている PageRank [1, 9] を計算するモジュールを実装して利用した．すなわち，他ページからの参照回数（被リンク数）に基づいてページの品質を評価する．$P(d)$ の値は，d が抽出されたページの PageRank 値である．ただし，検索のために集めた膨大な数のページ集合に対して現実的な時間で計算が終了するように，種々の近似計算を行っている．

タグ構造に基づいて用語説明らしさを確率的に定量化することは比較的困難であるため，経験的な値によってレイアウトモデルを実現した．具体的には，対象用語が HTML タグによって見出し語やリンクされている段落には 1 を与え，それ以外の段落には 0.5 を与える．

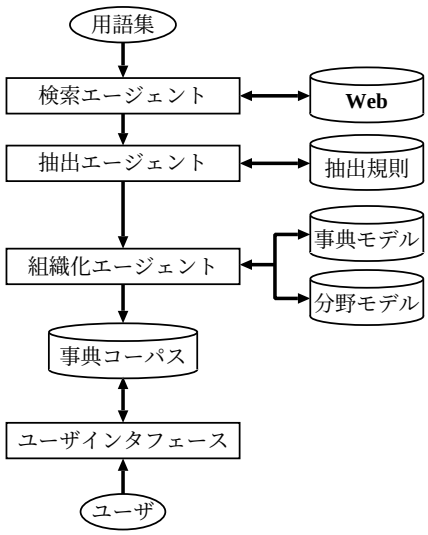

図 1: 事典システムの構成

(3) 事典コーパスの検索

構築したコーパスを既存の（オンライン）事典と同じように活用するためには，必要な項目を容易に検索するための仕組みが必要である．そこで，ユーザインタフェースを開発した．図 2 は，Web ブラウザ上で動作する検索用ユーザインタフェース画面である．事典コーパスは 2 通りの方法で検索することが可能である．

まず「用語検索」は，見出し語を入力して，その説明を取得する検索方法である．図 3 は「ウォークスルー」

[*2] http://chasen.aist-nara.ac.jp/
[*3] http://www.nova.co.jp/
[*4] http://www.google.com/

を入力したときの検索結果画面である．下線が付いた行は，説明を抽出した元ページのタイトルであり，それぞれの下に抽出された説明が表示されている（既存の検索エンジンの検索結果におけるページタイトルと内容サマリの一覧表示に似ている）．

図 3 では，コンピュータ分野の説明 3 件と機械工学に関する説明 1 件が画面内に表示されている．このとき，図 2 の「用語の種別」や「分野の詳細設定」によって，出力する内容をユーザが任意に選択することができる．

別の検索方法として「概念検索」があり，いわゆる「逆引き」が可能である．すなわち，検索質問に関連する用語説明を探索し，その見出し語を取得することができる．具体的には，用語説明を個別の文書と見なして単語単位で索引付けし，確率型の検索手法 [10] によって，入力された検索質問（単語，句，文など）に適合する用語を検索する．

その結果，コーパスに収録されている見出し語が分からない場合や，入力すべき語が思いつかない場合に比較的自由な入力ができるようになる．例えば「大画面の薄型ディスプレイ」を入力すると図 4 に示すように「PDP（プラズマディスプレイパネル）」や「プラズマディスプレイ」などの関連語が検索される．ここで用語を選択（クリック）すると説明を見ることができる．図 5 は「PDP」を選択した場合の表示内容である．すなわち，概念検索機能を使うと，概念的な要求によって具体的な見出し語を特定し，さらに，その語に関する説明を調べることができる．

また，ユーザは見出し語を入力したつもりでも，その語がコーパス中に見出し語として収録されていない場合には，入力語を含む用語説明が検索対象になるので，一種の「関連語検索」としても機能する．

(4) 大規模事典コーパスの構築

既存の事典を越えるためには，筆者らは，少なくとも 10 万語単位のコーパスを実際に構築し，種々の評価を通してシステムやコーパスを改善することが必要であると考えている．事実，言語モデルの学習に利用した「世界大百科事典」には既に約 8 万語が収録されている．

複数の資源から専門用語を中心に様々な用語を収集してコーパスの構築を行った結果，見出し語数は 208,962 に到達した．確率スコアに対する閾値を経験的に設定し，閾値以上の用語説明と分野のみ残したところ，用語あたりの説明数は平均 3.1 件，分野数は平均 2.0 件となった．

用語集が与えられれば，バッチ処理によってコーパスを自動的に拡張できることは当然のことに思えるかもしれない．しかし，主に計算機資源に起因する物理的・時間的制約があり，実際に構築するまでは，その実現可能性について知ることは不可能であった．

当該コーパスは人間が事典的に利用するだけでなく，例えば，事典情報を用いたシソーラスの構築 [6] や質問応答システム [8] のような計算機用辞書としての応用も可能である．通常これらの計算機処理は事典情報の不足が障害になるものの，その障害は今後解消されることが期待できる．

4 評価実験

(1) 実験方法

本事業で構築したコーパスは，既存の事典を越える数の見出し語を収録しており，量的問題は解消された．他方において，質的問題がどの程度解消されたかについても調査する必要がある．そこで，人間判定者がコーパス中の用語説明を評価した．しかし，構築したコーパス全件を評価することは非常に高価なので，対象用語を限定した．具体的には，情報処理技術者試験[*5]に頻出する語を集めた用語辞典 [12] に収録されている 2,226 語（見出し語と本文中の重要語として巻末の索引に収録されている語）を対象に実験を行った．

表 1: 「何を正解とするか」に関する 4 つの基準

判定者の意見 \ 度合	完全正解のみ	部分的正解も含む
一致	A	B
不一致も含む	C	D

対象用語はコンピュータ分野に関するものが中心であるものの「限界利益」や「OJT（on-the-job training）」などの他分野の用語や「エントロピー」のように複数分野で意味が異なる多義語も含まれていた．

上記 2,226 語を事典システムに入力した結果，2,080 語を含む事典コーパスが構築された（33 語は検索，113 語は抽出の過程で結果が得られなかった）．

本来ならば，事典システムにおける「検索」「抽出」「組織化」の全ての処理を評価するべきである．しかし，今回の実験では組織化だけを評価対象にした．

そこで，抽出処理で得られた段落をソートせずに提示し，判定者に用語説明として正しさと関連する分野の判定を依頼した．ただし，判定コストを削減するために，検索処理において各用語に対して最大 500 ページまでしか出力しなかった．4 人の判定者（本稿著者以外の大学卒業者）が分担して抽出結果を全て判定した．人間の判定は主観的になる可能性があるため，個々の用語に対して複数の判定者が判定を行うように分担方法を工夫した．具体的には，4 人を 2 つのグループに分けて，それぞれのグループが個別に 2,080 語を全て判定した（各グループ内では，2 人の判定者が 2,080 語を分担）．すなわち，一つの用語に対して 2 人が個別に判定結果を出した．一般に，2 人の判定者の結果が一致した判定ほど信頼性は高い．

個々の用語説明に対して「正しい」「部分的に正しい」「正しくない」の 3 段階で判定を行った．それ自身は完全な説明でなくても，説明の一部になりうる場合は「部分的に正しい」と判定した．対象用語について概要を知りたいユーザにとっては部分的に正しい用語説明でも十分である点に注意を要する．

以上より「何を真の正解とするか」という点に関して，表 1 に示す 4 つの異なる基準が成立する．

さらに，判定者に（部分的に）正しい用語説明に対して，関連する分野を一つ以上選択してもらった．候補となる分野は，分野モデル構築に利用した 19 分野（(4) 節参照）と「それ以外」の合計 20 分野である．約半分の用語説明はコンピュータ分野に関するものであった．しかし，残りの半分はコンピュータ以外の分野にほぼ一様に分布していた．

組織化による分野分類およびソートの目的は，通常の情報検索のように正しい情報を網羅的かつ高順位で出力することではない．類似した用語説明をいくつも見なくても，一つの用語説明だけでユーザが満足することは十分にありうる．そこで，再現率は評価尺度として利用しなかった．その代わりに，正しい用語説明が最初に見つかる順位を評価尺度とした．そして，対象となった全ての用語と分野に対する平均順位を計算した．

実際のコーパス構築では，確率スコアに対する閾値を設定し，スコアが小さい分野や用語説明は削除する．しかし，今回の評価では正しい用語がどのように順位付けされるかに焦点を当てたので，閾値による削除は行わず，

[*5] http://www.jitec.jipdec.or.jp/

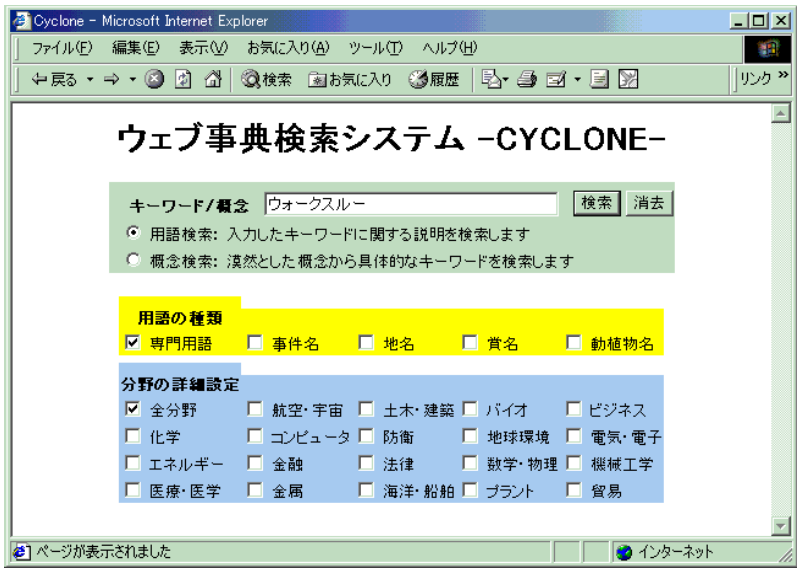

図 2: 事典システムのユーザインタフェース

図 3: 入力語「ウォークスルー」に関する説明

出力された用語説明全てを分野ごとに確率スコアに基づいてソートした．

さらに，単一の用語説明が複数分野に関連する可能性があることや，分野分類の精度が 100%でないことを考慮して，各用語説明を $P(c|d)$ の値に基づいて上位 5 件の分野に分類した．

(2) 実験結果

組織化処理における 4 つのモデル，すなわち信頼性（R），レイアウト（S），言語（L），ドメイン（D）の有効性を個別に調べるために，異なるモデルの組合せに対して，正解が最初に見つかる平均順位を比較評価した．ここでは，R，RS，RSL，RSLD の 4 通りについてのみ結果を示す．

表 2 に，異なる正解基準 A〜D（表 1 参照）に対する評価結果を示す．ここで「語義」は用語と分野の組合せを指す．また「用語数」と「語義数」は正解が最低 1 件存在した用語数と語義数である．判定者の一致度や正しさの度合を上げるほど対象用語や語義が少なった．しかし，いずれの基準においても，使用するモデルの種類を増やすほど，平均順位が向上した．この結果から，各モデルはそれぞれ事典コーパスの質を高める効果があったことが分かる．

分野モデルを使わない場合と使う場合は出力の形式が異なるため，同じ基準で評価するためには扱いに注意が必要である．分野モデルを使わない手法の場合，単一の用語に対する用語説明は全て一つのリストとしてソートされている．それに対して，分野モデルを使用する手法

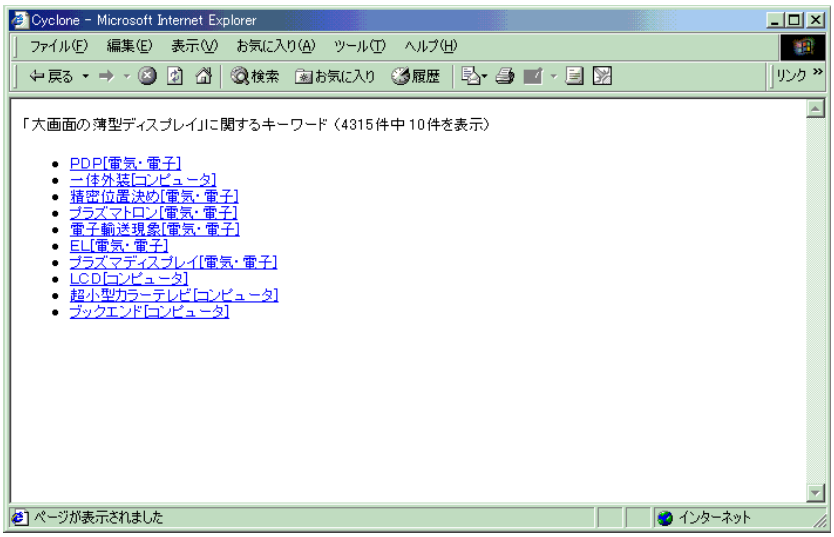

図 4: 概念検索の例

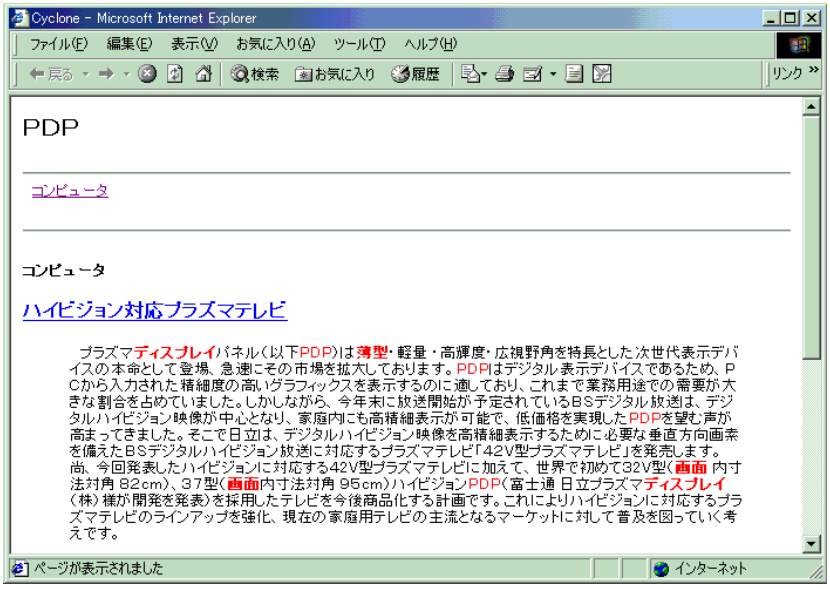

図 5: 「PDP」に関する説明

では上位 5 件以外の分野に対しては用語説明が割り当てられないため，分野モデルを利用しない手法よりもリストが短くなる．その結果，平均順位を同じ基準で比較することができなくなってしまう．この問題を避けるために，上位 5 件の分野以外については $P(c|d)$ の値を限りなく 0 に近い正数にして，各分野リストの末尾に加えた．以上をまとめると，表 2 において，異なる手法の平均順位は比較可能な値である．

表 2 を見ると，分野モデルによる平均順位の向上が比較的小さい．分野モデルは本来，複数の分野に対応する多義語に対処するために導入されたモデルである．そこで，多義語だけを対象にして平均順位の比較を行った結果を表 3 に示す．この結果では，表 2 の場合に比べて RSL と RSLD の違いが顕著になり，分野モデルがより効果的であったことが分かる．

さらに，上位 10 件に順位付けられた用語説明の分布を調べた結果，全体の 80% の語義は上位 10 件以内に正しい用語説明が見つかった．図 3 に示したように，各用語説明は短い段落で構成されているので，上位 10 件を閲覧することは比較的容易である．そこで実用上の問題はそれほど大きくない．

以上の実験結果より，組織化処理に利用したモデルは個別に効果的であり，分野モデルは多義語に対して特に効果的であることが分かった．

他方において，分野モデルを利用しても語義を区別できない場合があった．図 3 において，最初の 3 件はいずれもコンピュータ分野の説明であるものの，最初の説明は「3 次元 CG による仮想空間内の移動」であり，残りの 2 件は「ソフトウェア開発における成果物の検証」に関する説明である．ただし，機械工学に分類された段落

表 2: 正しい用語説明が最初に見つかる平均順位（R:信頼性, S:レイアウト, L:言語, D:ドメイン）

			使用したモデル			
正解の基準	用語数	語義数	R	RS	RSL	RSLD
A	634	724	39.9	33.5	23.2	22.8
B	952	1,112	37.0	30.0	18.0	17.0
C	1,539	2,212	41.7	35.3	21.7	19.4
D	1,678	2,633	37.8	33.3	20.0	17.4

表 3: 正しい用語説明が最初に見つかる平均順位（多義語のみ対象）

			使用したモデル			
正解の基準	用語数	語義数	R	RS	RSL	RSLD
A	36	74	59.7	52.7	31.8	28.4
B	47	127	67.4	58.5	35.2	28.0
C	476	1,112	56.4	50.6	31.2	27.0
D	501	1,376	53.0	49.3	30.5	25.8

は，車の種類（運転席から車両後部まで歩いて移動できるタイプの車）について説明しており，コンピュータ分野とは区別された．

5 モニタ調査に基づく事典システムの評価

(1) 調査目的

構築した事典コーパスを評価するために，不特定多数の被験者によるモニタ調査を実施した．調査の第一目的は，事典システムとして整備した内容が，一般ユーザが事典で検索したくなるような「意味の分からない用語」を含んでいるか，抽出された用語説明は適切かどうか，を検証することである．第二の目的は，事典システムが，一般的な検索システムと比較して用語を検索する用途に向いているかどうかを検証することである．この調査を行うため，一般的な検索サービスと同様に Web ブラウザに入出力インタフェースを実装した．

(2) 調査方法

本調査は「検索サービスモニタ募集」のメールに返答したインターネット調査会社のモニタ会員に対して行った．具体的には，20〜59 歳の幅広い年齢層の被験者を対象に 1 週間かけて調査を行った．

(3) 調査内容

各被験者は，事典システムに自由な用語（テレビ，新聞，雑誌などで見つけた言葉など）を入力し，提示された結果が「分かりやすい」かどうかを主観で評価した．

どの程度の語数を準備すれば実用に耐えうるかを調べるために，語彙の規模を 5 万，10 万，15 万，20 万と段階的に変化させたコーパスを個別に用意し，それぞれ異なる約 250 名（計約 1,000 名）に 2 語ずつ入力させた．

調査に用いた質問票の具体的な内容は以下の通りである．まず，全ての被験者に次の質問を行った．

- 普段使っている検索サービスは何か?

さらに，以下の質問を用語ごとに別々に回答させた．

- テレビ・新聞・雑誌などで見つけた意味のよく分からない用語は何か?
- 事典システムでその用語の説明が表示されたか?
  - 表示された場合は何番目の説明で分かったか?
- 関連語検索を利用して用語の意味が分かったか?
  - 表示された場合は何番目の説明で分かったか?
- その用語について普段から利用している検索サービスで検索してみて，事典システムの結果と比較してどちらが分かりやすかったか?

(4) 調査結果

調査は，語彙サイズが異なる 4 つのコーパスごとに別々の被験者を募って行った．目標の回収数である 250 名を達成するために，各コーパスごとに，620 名ずつ依頼のメールを送った．最終的な回収数は表 4 のように全体で 1,038 名分，回収率は 41.9%になり，当初の目的はほぼ達成された．回収できた回答者の性別比は男性 44.4%，女性 55.6%であった．年代と職業の分布をそれぞれ表 5 と表 6 に示す．

表 4: 回収数

語彙サイズ	5 万語	10 万語	15 万語	20 万語	合計
回収数	257	261	244	276	1,038

表 5: 被験者の年代に関する分布

年代	20	30	40	50
比率（%）	37.2	42.3	16.5	4.0

表 6: 被験者の職業に関する分布

職業	比率（%）
会社員	41.9
専業主婦	22.3
学生	9.3
パート・アルバイト	7.6
自営業	5.8
公務員	3.9
専門職 (医師・弁護士等)	1.4
教職	1.4
無職	2.6
その他	3.8

モニタが普段使っている検索サービスの割合を図 6 に示す．Yahoo はキーワード検索に Google を利用しているので，全体の約 75%は事実上 Google を利用している．

まず，本事典システムの用語検索（見出し語を入力する検索方法）の性能についてまとめる．語彙サイズごとの性能を図 7 に示す．語数が増えるにつれて性能が向上していることが分かる．20 万語コーパスの場合は，約 40%の被験者が用語検索によって出力された説明で入力した用語の意味が分かったと回答した．

次に，関連語検索まで含んだ結果を図 8 に示す．用語検索と関連語検索をあわせた全体的な性能も，語数が増えるにつれて向上したことが分かる．20 万語コーパスの場合は，約 60%の被験者が事典システムの出力によって，入力した用語の意味が分かったと回答した．

さらに，普段使っている検索サービスとの比較を行った結果を図 9 に示す．まず，コーパスの規模に拘らず，事典システムでも検索サービスでも分からなかった語の割合は約 15%であった．この数字は，現行の検索サービスの限界点を示している．

また，語数が増えるごとに（普段使っている検索サービスと比較した場合の）本事典システムに対する印象が良くなっている．20 万語コーパスの場合は，事典システムの方が分かりやすいという回答が 36.0%であったのに対し，検索サービスの方が分かりやすいという回答が 33.6%であり，ほぼ同等であった．

事典システムの性能をより詳細に検討するため，20 万語コーパスの検索結果について，事典システムと検索サービスのどちらの方が分かりやすかったかという比較

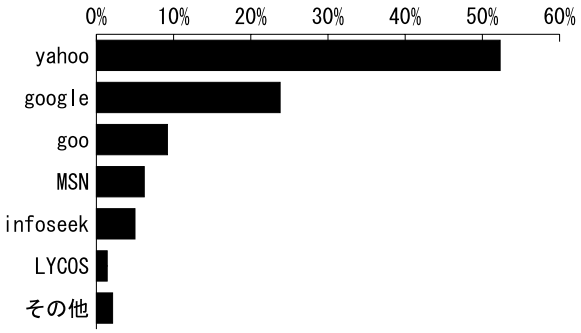

図 6: 被験者が普段利用している検索システム

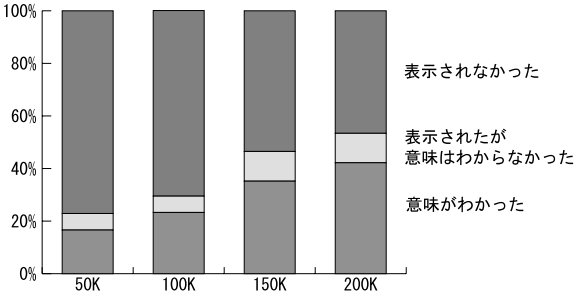

図 7: 用語検索の評価結果

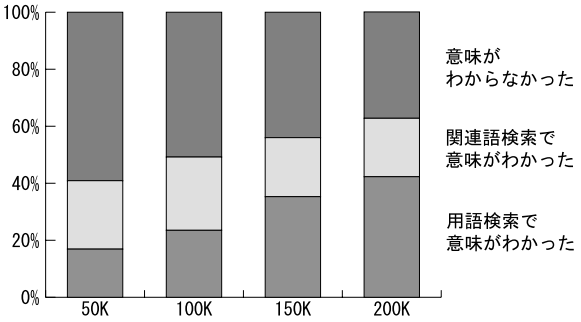

図 8: 関連語検索まで含めた場合の事典システムの評価結果

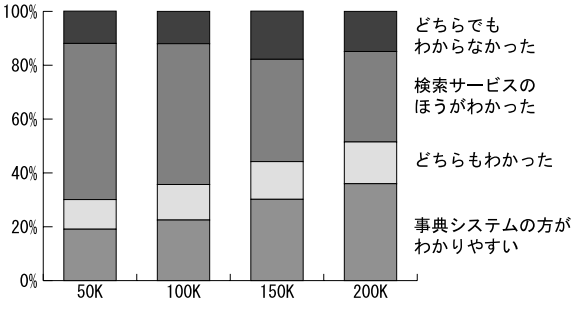

図 9: 事典システムとモニタが普段利用している検索サービスの比較

結果と，事典システムを用いた場合の検索結果（用語検索，関連語検索で分かったか，どちらでも分からなかったか）との関係を調べた．その結果を図 10 に示す．この図において，横軸は「事典システムと検索サービスのどちらの方が分かりやすかったか」に対する回答である．縦軸は，事典システムを用いた場合の検索結果（用語検索，関連語検索で分かった，どちらも分からなかった）を語数で示したものである．

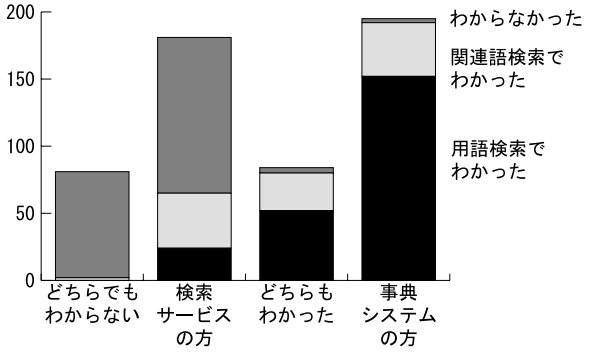

図 10: 比較結果と検索結果の関係

図 10 より「普段使っている検索サービスの方が分かりやすかった」と回答した被験者（左から 2 番目）のうち，約 60%は「事典システムでは意味が分からなかった」と回答している．しかし「事典システムの方が分かりやすかった」と回答した被験者（最右）の約 80%は用語検索で意味が分かったと回答している．したがって，カバーする用語の数を 20 万よりさらに増やして用語検索によるヒット件数が増えれば，既存の検索サービスよりも有用なサービスになることが期待できる．

最後に，被験者が入力した用語に関して詳細に分析した．被験者が入力した語は異なりで 1,124 語あった．上位 10 語の頻度を表 7 に示す．

表 7: 被験者が入力した語の例と入力された頻度

用語	頻度
ペイオフ	151
政策秘書	43
ブロードバンド	37
外形標準課税	33
ワークシェアリング	28
更迭	26
ユビキタス	17
証人喚問	15
ADSL	13

表 7 に示した語のうち，上位 5 件に関して，さらに詳細を分析した．

「ペイオフ」はコーパスの規模によらず，用語検索で正しい説明を得られたという回答がほとんどであった．回答者が満足した説明の例（本システムでは，用語検索の第一候補として出力された）を以下に引用する[*6]．

> ペイオフとは，金融機関の経営が破綻して預貯金などの払い戻しを停止または停止する恐れが生じた場合，金融機関が貯金保険機構や預金保険機構に積み立てている保険金で，破綻した金融機関に代わって，預貯金者 1 人につ

[*6] http://www.kahoku.co.jp/NEWS/2001/11/20011107J_12.HTM

> き元本１千万円とその利息を限度として払い戻しを行う制度です．１千万円を超える部分の元本とその利息については，破綻した金融機関の清算後の資産で弁済を見込める割合に応じて預貯金者に配分されます．現在，特例措置として「ペイオフ」は凍結されており，預貯金の金額が保護されていますが，平成１４年４月からこの特例措置が解除され,「ペイオフ」が解禁される見込みです．

「政策秘書」は用語検索については語彙の規模に拘らず検索できなかったものの，ほとんどの被験者が関連語検索によって意味が分かったと回答した．関連語検索によって表示された説明の例を以下に引用する[*7]．

> 平成５年度に,「政治改革」の気運を受けて誕生した国会議員政策担当秘書は，国会議員が政策や法律を立案する際に立法調査などの業務を行い，議員の活動を補佐するために設けられた資格です．アメリカでは立法調査官として，政策や法律の立案に大きな役割を果たしています．議員立法の活性化が叫ばれる中，大きな活躍が期待されています．

「ブロードバンド」はコーパスの規模にかかわらず用語検索では検索できず，関連語検索を使っても適切な説明は得られなかった．また「外形標準課税」と「ワークシェアリング」に関しては，5 万語コーパスでは用語検索で説明を得られていなかったものの，10 万語以上では，ほとんどの被験者が用語検索で説明を得られたと回答した．

(5) モニタ調査のまとめ

開発した事典システムの性能を評価するために，Web を通じて約 1,000 名の被験者による検索実験を行った．この結果，コーパスの語彙サイズ（見出し語数）を大きくすると，総じて性能が向上することが分かった．被験者の入力した用語に対して，用語の意味が分かる説明を出力することができた語の割合は，5 万語の場合には約 40%だったのに対して，20 万語の場合には約 60%であった．また，20 万語コーパスでは，通常の検索サービスと同等の評価を得ることができた．本事典システムに対する不満は，主に収録されている見出し語数の不足に起因するものなので，今後見出し語数を増やせば（例えば 50 万語），既存の検索サービスに匹敵するか，もしくはそれらを越えられる可能性がある．

6 本事業のまとめと今後の展望

本事業は，World Wide Web から大百科事典を自動構築することを目的とし，Web から用語説明を収集して高品質の説明を分野ごとに整理・組織化するためのソフトウェアエージェントを開発した．さらに，開発したエージェントを用いて収録語数約 20 万語の巨大な事典を構築した．事典を活用するためのインタフェースをあわせて開発し，以上を総括して事典システムを完成させた．

複数の判定者による評価実験によって，個々の処理（エージェント）が事典の品質を高めることに有効であることが分かった．また，不特定多数の被験者によるモニタ調査を実施した結果，既存の検索エンジンよりも，本システムは事典的な使用に適していることが分かった．

本事典システムを Web 上のサービスとして今後さらに発展させるためには，収録語数の拡充が必要である．また，長期に渡って継続的に運用するためには，新語を自動的に発見して収録語として追加する機能や既収録語に対する事典情報を更新する機能についても検討し，開発を行う必要がある．

[*7] http://www.lec-jp.com/learning/k7/info_002.html

7 参加企業及び機関

（財）日本産業技術振興協会
（プロジェクト実施管理組織）

8 謝辞

本事業の機会を与えて下さった大黒晶議プロジェクトマネージャーに感謝致します．

9 参考文献

[1] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, Vol. 30, No. 1–7, pp. 107–117, 1998.

[2] Philip Clarkson and Ronald Rosenfeld. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of EuroSpeech'97*, pp. 2707–2710, 1997.

[3] Atsushi Fujii and Tetsuya Ishikawa. Utilizing the World Wide Web as an encyclopedia: Extracting term descriptions from semi-structured texts. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 488–495, 2000.

[4] Atsushi Fujii and Tetsuya Ishikawa. Organizing encyclopedic knowledge based on the Web and its application to question answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 196–203, 2001.

[5] Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. Producing a large-scale encyclopedic corpus over the Web. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 1737–1740, 2002.

[6] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pp. 539–545, 1992.

[7] Makoto Iwayama and Takenobu Tokunaga. A probabilistic model for text categorization: Based on a single random variable with multiple values. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pp. 162–167, 1994.

[8] Julian Kupiec. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 181–189, 1993.

[9] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the Web, 1998. http://www-db.stanford.edu/~backrub/pageranksub.ps.

[10] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[11] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189–196, 1995.

[12] 藤本喜弘（編）. 第二種・シスアド情報処理用語辞典. 経林書房, 1998.

[13] 藤井敦, 伊藤克亘, 石川徹也. WWW は百科事典として使えるか? –大規模コーパスの構築–. 情報処理学会研究報告 2002-NL-149, pp. 7–14, 2002.

[14] 藤井敦, 石川徹也. World Wide Web を用いた事典知識情報の抽出と組織化. 電子情報通信学会論文誌, Vol. J85-D-II, No. 2, pp. 300–307, 2002.