事典的Web検索サイトの構築

藤井 敦 †,†††　伊藤克亘 ††,†††　秋葉友良 ††
† 筑波大学
†† 産業技術総合研究所
††† 科技団 CREST
fujii@slis.tsukuba.ac.jp

1 はじめに

近年，World Wide Web を一種の辞書のように使って，知らない言葉や事柄を調べることが一般的になってきた．Web には既存の辞典や事典に載っていない多種多様な情報が存在するからである．Web が流行りはじめた当初に比べれば，検索エンジンの性能は向上し，目的の情報が簡単に見つかることも多くなった．しかし，検索要求によっては，依然として何を入力すればいいのか分からない場合や，膨大な検索結果から欲しい情報をどうやって選択すればよいか分からない場合がある．また，Web には統制がないため，誤字，誤解，虚偽といった低品質情報を排除する必要がある．

筆者らは Web を事典的に利用することを目的として，Web ページ群から言葉に関する説明情報を収集する手法 [1, 5, 6] を提案した．本手法は，Web に含まれる良質な説明情報を選択的に抽出し，専門分野に基づいて分類することで事典コーパスを自動構築する．本手法を用いて約 20 万語の見出し語を含む大規模な事典コーパスを構築した [5]．その結果，Web ブラウザを介して様々な用語に関する説明を分野単位に検索することが可能になった．すなわち，原理的には事典情報を対象とした専門検索サイトの原型を実現したことになる．

しかし，当サイトを継続的に運用し，また「使える」サイトとして位置付けるためには様々な問題を克服しなければならない．こうした背景を踏まえ，本稿は事典的 Web 検索サイトの構築に関する種々の手法を提案する．

2 検索サイトの主旨

検索サイトを「使える」ものにするためには，サーバ耐久性の向上，ページのリンク切れに対応するキャッシュの充実といったハードウェアに関する側面から，コンテンツ（事典コーパス）の品質向上，検索インタフェースの利便性向上といったソフトウェアに関する側面まで幅広い工夫の余地がある．本稿ではソフトウェアに関する側面からの工夫に焦点を当てる．

本サイトの主旨は，とにかくユーザを飽きさせないことである．ネットサーフィンが流行る主な理由は，マウスのクリックによってハイパーリンクを辿るだけで様々な情報を簡単に取得できる点にある．言い換えれば，それ以上先に進めないような行き止まりに陥ると，ユーザの不満は大きくなる．これは，本事典検索サイトのユーザにも当てはまる．すなわち，

- どんな見出し語を入力すればよいか分からない
- 入力した語が見出し語として登録されていないために何も検索されない
- 検索された説明が分かりにくい，もしくは説明になっていない

などの理由で検索行動の中断を余儀なくされた場合，ユーザは検索サイトの利用をやめる．事実，約 1,000 人の被験者を対象にした調査の結果，見出し語のヒット率がユーザの満足度と強く関連することが分かった [5]．

以上をまとめると，ユーザの入力に対して常に何らかの（意味のある）応答をして，またユーザが目的の説明を見つけた場合でも，次の検索へ自然に誘導するような仕組みが必要になる．

3 検索サイトの機能

図 1 に基づいて検索サイトの機能について説明する．簡単のため，事典コーパスを構築するオフライン処理と，ユーザが検索するオンライン処理に分けて説明する．

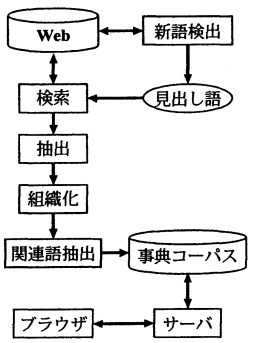

図 1: 事典的 Web 検索サイトの概要

オフライン処理では，まず「新語検出」によって見出し語の候補をWebから自動的に収集する．次に，各候補に対して「検索」「抽出」「組織化」を順番に実行し，説明を専門分野ごとに分類する．そこで「パイプライン（方式/油送管）」のように分野によって意味が異なる多義語の説明を区別することができる．全ての候補に対して必ずしも説明が得られる訳ではない点に注意を要する．

検索処理は，見出し語を検索語としてページを検索する．抽出処理は，HTMLタグを用いて見出し語に関する説明を段落単位に抽出する．組織化処理は，a) 特定分野への関連度，b) 説明らしい言語表現を含むかどうか，c) 説明らしいHTMLレイアウトかどうか，d) ページの信頼度という4つの尺度を統合したスコアを計算して，その値に基づいて段落を分野に分類し，ソートする．

最後に「関連語抽出」によって，見出し語を特徴付ける語を取得する．これらの語は，オンライン検索時にユーザの情報要求を絞り込むために利用する．

オンライン処理では，Webブラウザを介して事典コーパスを検索する．本稿執筆現在，見出し語数は約60万である．図2は，見出し語「ウォークスルー」で検索した結果である．この用語には「攻略法」「仮想空間の移動」「ソフトウェア開発工程の検証」「車種」「ゲート型の金属探知機」等の異なる意味があり，図2には最初の2つの意味に関する説明が表示されている．

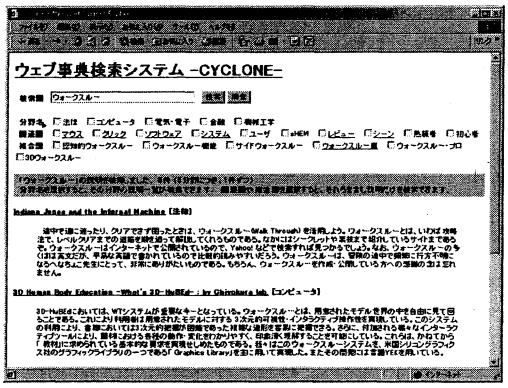

図 2: 入力語「ウォークスルー」に対する検索結果

これら多数の意味を全て自動的に区別することは困難である．そこで「ウォークスルー」に関連する分野や語を提示し，ユーザが選択した分野や語に関する説明だけを選択的に提示することでユーザの情報要求を絞り込む．これは，雑多な説明の中でユーザの検索行動を中断させないための工夫である．例えば「マウス」という関連語を選択すれば「仮想空間の移動」に関する説明が提示される．他方で「ウォークスルー車」を選択すれば「車種」に関する説明が提示される．さらに，提示された関連語が見出し語として登録されていればリンクが張られる．ユーザはネットサーフィンと同じ要領でリンクを辿るだけで，関連する語の説明を次々と簡単に調べることができる．

ユーザが入力した見出し語が事典コーパスに無い場合は，種々の補完機能によって関連する情報を提示する．具体的には，前方・後方・部分一致を用いて文字列として類似する語を提示する．また，同じ英訳に対応する語を類義語として検出する．例えば「レイテンシー（latency）」が見出し語になければ「待ち時間」を代わりに提示する．

漠然とした要求はあるものの，何を見出し語とすればよいか分からないユーザ（場面）も想定している．例えば「電子メールに感染するもの」と入力すれば「コンピュータウイルス」や「マクロウイルス」のような見出し語を提示する．具体的には，説明情報に対する全文検索を行うことで，事典の逆引き機能を実装している．

入力された語が見出し語として存在しない場合は，オフライン処理における一連の処理を行って，説明情報を動的に生成することも可能である．しかし，1語の処理に数分を要するため，現実的な方法ではない．

今回新たに追加した「関連語抽出」と「新語検出」について4章と5章でそれぞれ説明する．

4 関連語抽出

関連語抽出の基本原理は，各用語の説明情報（段落）に頻出する語を検出することである．ここで，適切な語を検出する処理と検出した語を評価する尺度が必要になる．そこで，まず段落を「茶筌」で形態素解析して，品詞情報に基づいて（複合）語を構成し，関連語の候補とする．具体的には，名詞，動詞連用形，未知語，記号の連続を語として抽出する．さらに，段落における出現頻度と抽出元の段落に対する組織化のスコア（3章参照）を統合して関連語をソートし，上位の関連語から優先的に提示する．すなわち，良質の説明によく現れる語が優先される．

実際は，見出し語を含むかどうかによって関連語を2種類に分類する．以降では，見出し語を含む語を「複合語」，含まない語を「関連語」と呼ぶ．前者は，例えば「ウォークスルー」に対する「ウォークスルー車」のような複合語である．多義語は，前後に連接する語によって語義を特定できる場合があるため，複合語による情報要求の絞り込みは有効である．

他方おいて，関連語とは，例えば「ウォークスルー」と共出現する「マウス」や「レビュー」などの語である．これらの語は複合語と同様に多義性を解消する目的のために有効である．さらに，同じ語義に関する説明であっても，説明の観点が変わる場合がある．関連語はこのような観点の違いを区別するためにも有効である．例えば「特許法」という見出し語に対して「特許料」「存続期間」のような関連語が提示される．これらの関連語は特許法の説明を異なる観点から調べるために有効である．

5 新語検出

5.1 新語検出の意義

本検索サイトで利用されている事典コーパスは，現状でも有用性が高い資源である．しかし，更新頻度が高いというWebの特徴を利用すれば，Webから新しい見出し語を継続的に取得して事典コーパスを更新し，有用性をさらに高めることができる．

従来，本サイトでは見出し語リストを人手で与えていた．しかし，事典コーパスの自動拡張を目的として，更新頻度が高いWebページや新聞記事などの文書群から新語を検出するモジュールを開発した．また，そのモジュールを用いて評価実験を行った．

5.2 「新語」の分類

本検索サイトにおける「新語」とは，事典コーパスに見出し語として登録されていない語である．新語は以下のように分類できる．

a. 既存の語であるにも拘らず，事典コーパスの見出し語として登録されていない語
b. 新しく作られた言葉（「ムネオハウス」など）
c. 異表記（「ら致」と「拉致」など）
d. 既存の語に対する新しい用法（後部座席から前方の座席に車内で移動できるようにする機能やその機能を持った乗用車を指す「ウォークスルー」など）

また，不特定多数の著者が執筆した不均質なWebページ群を対象にする場合は，文書の不備（誤字，誤殖，脱字など）に起因して，未知語問題が生じることもある．

このように様々な側面のある新語のうち，本稿では，表記上の「新語」だけを検出の対象とする．そこで，上記のうち，a，b，cだけを対象とする．dは，既収録語に対して，図1における「検索」以降の処理を定期的に行うことで，説明情報を更新することで対処する．

日本語では語の検出（認定）自体が困難であり，様々な手法が提案されている [2]．特に事典見出し語の語構成を考えた場合は，語は単独の形態素で構成される場合だけでなく，複合名詞に代表される複合語や複雑な構造を持つフレーズの場合もある（人工知能分野における「説明に基づく学習」など）．また，近年はbに分類される新語として「ちょぼら（ちょっとしたボランティア）」「モーニング娘。（句点が語の一部として含まれている）」「１０９（デパート名）」のような特殊な語が増えており，新語検出を一層困難にしている．

5.3 新語検出手法

本検索サイトでは，語の説明を取得するために，1日ごと，1週間ごと，1ヶ月ごとなど期間を定めて，その間に新たに収集した文書集合から新語を検出する．継続的に語彙を拡張しながら検出することが前提となるため，既知語のリストを参照しながら，未登録の新語を検出する．そこで，既知語リストを用いて新語を検出する手法を提案する．検出の手順は以下の通りである．

1. 新語検出の対象となる文書を形態素解析する．前処理として，対象がHTMLファイルなどの構造化文書の場合はタグを削除し，さらに半角文字を全角文字に変換して統一する．
2. 形態素列から，名詞（数字を含む），未知語，記号（「・」など）の連続を新語候補として抽出する．

   この処理では「説明に基づく学習」のようなフレーズは抽出されない．他方で「03-3333-3333」のような電話番号や単なる数字（列）が抽出されるため，新語候補数は膨大になる．
3. 対象とする文書全体の抽出結果から，新語候補の頻度リストを作成する．
4. 頻度リストから，既知語リストに登録されている語を削除する．この段階で大半の語は削除される．
5. 4で作成した頻度リストから，低頻度語を削除する．ただし，低頻度語も既知語リストには追加する．そこで，ある期間内に低頻度で出現した形態素列は，当該期間に関しては，低頻度であるために削除され，当該期間以降は既知語リストに登録されているために新語候補として検出されない．

   その結果，連絡先の電話番号や数字を削除でき，さらに，誤字，脱字，形態素解析誤り（の一部）も削除できる．

意味のない新語候補は，後続の処理で説明情報が得られない可能性が高いため，その段階で淘汰される．そこで，本新語検出手法は，精度よりも網羅性を重視している．また，多様な見出し語を検出するため，字種や長さに関する制限は加えない．他方において，不要な候補を増加させると，後段の処理効率を低下させてしまうため，実際の運用においてはコストを考慮する必要がある．

5.4 評価実験

新語とは，既存語に対する相対的な存在であるため，何が新語であるかという絶対的な基準はない．そこで，新語検出の性能評価は容易ではない．用語抽出の評価用コレクションとして，例えば，論文抄録を対象にしたNTCIR [3] がある．しかし，本検索サイトでは，更新頻度が高い新聞記事（サイト）などから新語を検出するため，新聞記事を対象にした評価が好ましい．毎日新聞CD-ROMには，記事検索のためのキーワードリストが記事ごとに付与されている．このキーワードリストを正解として利用した．2000年版の統計を表1に示す．

ただし，キーワードリストには「との中期的見通し」のように前後が助詞である表現も多数含まれている．さらに「中期」「中期的見通し」のように一部が重複した

表 1: 毎日新聞 CD-ROM の形態素/キーワード数

異なりキーワード数/記事	117
異なり形態素数/記事	357
異なりキーワード数/年	1,852K

キーワードも多数含まれている．理想的には，事典の見出し語としてふさわしい語だけを選択して正解とすることが好ましい．しかし，判定のコストが高いことから，今回はキーワードリストに含まれる全てのキーワードを正解と見なした．

形態素解析には「茶筌」を用いた．既知語のリストは，毎日新聞 CD-ROM 1991 年版から 1999 年版までの東京版の記事を用いて作成した．毎日新聞 CD-ROM 2000 年版の記事を用いて月単位で新語を検出した．検出時には月単位で 2 回以下の出現頻度のものは削除した．また，検出した新語は順次既知語のリストに追加して実験した．最終的に，既知語リストに含まれる単語数は 4,593K から 5,010K まで増えた．

実験結果を表 2 に示す．ここでは，検出された新語候補数（検出数），キーワードリストに含まれた候補数（一致数），両者の比率（精度），正解だった新語のうち，当該月における頻度が最も高かった語（最頻出語）とその度数を示している．

検出される語を既知語リストに追加したにも拘らず，毎月 600 語以上の新語が検出され，月平均 831 の新語が検出された．ただし，時期によって検出数に変動が生じた．6 月は選挙のために立候補者名が記事に多数出現した．また，選挙の前後で議員の役職などが変わるため，○○幹事長の○○の部分が置換された語が多く出現し，かつ以前の幹事長は△△前幹事長となるために新語候補数が増加した．

語数の増加には結び付かなかったものの，11 月には，いわゆる「加藤の乱」があり，役職，所属する党，所属する派閥が変わる議員が多かった．12 月は 2000 年を振り返るための数字（例えば，国債発行残高の「１兆３５６８億円」）が多数使われた．

表 2: 新語検出の実験結果

月	検出数	一致数	精度	最頻出語	度数
1	675	575	0.852	カルマパ１７世	68
2	627	534	0.852	小林本部長	68
3	788	676	0.858	アナリエ	27
4	788	658	0.835	森首相	253
5	704	593	0.842	陳総統	46
6	1226	1019	0.831	モナザイト	154
7	855	717	0.839	２，１１３	354
8	898	770	0.857	倉容疑者	32
9	941	791	0.841	萩崎容疑者	58
10	936	765	0.817	コシュトゥニツァ	47
11	780	670	0.859	手集計	181
12	754	584	0.775	ブッシュ次期大統領	47
平均	831	696	0.838	—	—

6 おわりに

Web を事典的に利用できる検索サイトの構築手法について説明した．特に，見出し語を自動更新する新語検出の機能と実験結果について詳説した．今後は a) 検出された新語のうち，どの程度の語に対して説明が取得できるか，b) 新語検出後どれくらい経過すれば説明を含むページが Web に出現するか，といった観点から総合的な評価を行う必要がある．また，事典コーパスを用いた質問応答 [4] などを応用して，本サイトを質問応答サイトに拡張する可能性についても検討する．

謝辞

本研究の一部は，IPA 未踏ソフトウェア創造事業によって行われました．プロジェクトマネージャーの喜連川優先生（東京大学）からは有益なコメントを頂きました．

参考文献

[1] 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.

[2] Kyo Kageura and Bin Umino. Methods of automatic term recognition: A review. *Terminology*, Vol. 3, No. 2, pp. 259–289, 1996.

[3] Kyo Kageura, Masaharu Yoshioka, Koichi Takeuchi, and Teruo Koyama. Overview of the TMREC tasks. In *Proceedings of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, p. 415, 1999.

[4] 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.

[5] 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.

[6] 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.