ヒューマンインタフェース 64－9
（1996. 1. 11）

WWWでの辞書引き方法の比較検討

伊藤修一† 梅村恭司‡

† sito@avenue.tutics.tut.ac.jp
‡ umemura@avenue.tutics.tut.ac.jp
豊橋技術科学大学 情報工学系 梅村研究室
〒441 豊橋市天伯町雲雀ヶ丘 1-1

WWWではその言語，専門分野とも多岐に渡っている．そのためリファレンスシステムはWWW上で非常に有用である．現在リファレンスシステムは多数市場に出回っている．そこで我々はその分類を行なった．また我々はProxyサーバを利用したWWW上でのリファレンスシステムAutorefを開発した．本システムによれば，ユーザはマウスクリック一つで単語情報を参照することができ，WWWブラウザの種類を問わずに使用することができる．この論文ではまず，リファレンスシステムの分類について述べている．そして，本システムの実現方法を記述し，我々の分類に基づき各種の辞書引き方法を比較検討している．

Comparison and Taxonomy about Dictionary Looking Up Methods on WWW

Shuichi Itoh† and Kyoji Umemura‡

† sito@avenue.tutics.tut.ac.jp
‡ umemura@avenue.tutics.tut.ac.jp
Umemura Laboratory, Department of Information and Computer Sciences,
Toyohashi University of Technology
1-1, Tempaku, Toyohashi, Aichi 441, Japan

Reference systems become more useful for World Wide Web, since there are multilingual informations and many special fields on WWW. Therefore, we have classified many kinds of referencing systems. We have also developed a reference system, called Autoref, for WWW. This system uses Proxy mechanism and can be used with any kind of browsers. User can get the meaning of a word with a mouse click. This paper describes classifications of reference systems at first. Then, We have explained the implementation of our system, and we have compared and examined dictionary looking up methods.

1 はじめに

WWW(World Wide Web)には多種多様な情報が存在している．今では，何かを知りたいと思った時に，Internet上にその答えがある場合がほとんどであろう．しかし，その情報はたいてい分散して存在しており，欲しい情報の全てを1つの情報源から入手できるとは限らない．また，WWWが世界中に網細状に浸透してきたことにより，それ全体が大きな百科辞典として機能する可能性も出てきた[1]．それはYahoo(http://www.yahoo.com/)のようなInternetのイエローページから辿ることができよう．しかし我々はWWWサーバ上にあるリファレンスシステムがまだ十分に機能していないと考える．このような現状であるため，現在WWW上で簡単に利用できるリファレンスシステムが必要とされている．幸いにもいくつかのリファレンスはCD-ROM化されており，そして今後なおリファレンスの電子化は進むだろう．

WWWではドキュメントによって言語や専門分野が変わる[2]．よって，英単語の辞書引きをサービスとして行なうシステムは英語を母国語としないユーザにはより有益になる．たとえ英語が母国語であったとしても，専門用語は内容を理解する上で障害となるだろう．またその場所の話題が何によるか，どの辞書を参照すべきかを決めることが難しいことが時々ある．このような状況では，高機能で多目的な辞書引きシステムを利用することがとりわけ有効であると言える．

2 辞書引き方法の種類

現在様々な種類の辞書引き(リファレンス)システムが市場に出回っている．英語の文章を読み易くするという意味では全てのシステムは同じコンセプトを持っている．これらのシステムのユーザインタフェースは様々であるので，我々はこれを分類整理してみることにした．

まず，リファレンスの方法について考えてみると，1つ目として入力した文章の意味をその文脈まで考慮して機械翻訳する翻訳型があり，2つ目として入力された単語の意味をそのまま答えとして返すリファレンス型がある．

次にWWWへの関与の形を考えてみる．1つ目はWWWブラウザの機能に密接にリンクしているシステムがあげられる(ブラウザ型)．2つ目として，ProxyとWWWブラウザの中間にシステムを置くタイプがある(Proxy型)．ブラウザから見た場合，このシステムはProxyサーバのような振舞いをするであろう．最後にWWWに限定せず，汎用性を求めたタイプ(汎用型)がある．これら3つのタイプがあると我々は考える．

表1: リファレンスシステムの分類

リファレンスの方法		WWWへの関与の形		
		汎用型	ブラウザ型	Proxy型
翻訳	ページ分け	J-London[1] LogoVista[2]	NetSurfer[4] Transpad[5]	PENSEE[6]
	埋め込み		NetSurfer Transpad	PENSEE
リファレンス	ページ分け			Autoref
	埋め込み			EtoJ_Proxy[7]
	リタイプ	Quick Viewer[3]	NetSurfer Transpad	

[1] J-London/EJ ··· (株)高電社
[2] LogoVista E to J ··· カテナ(株)
[3] Quick Viewer ··· (株)HAL研究所
[4] Netsurfer/ej ··· (株)ノヴァ
[5] Transpad for Windows ··· 亀島産業(株)
[6] PENSEE for Internet ··· 沖ソフトウェア(株)
[7] EtoJ_Proxy ··· 文献[3]

さらにブラウジングの方法もいろいろある．まず1つ目は単語を入力してその意味を得るタイプ(リタイプ型)．これは他の方法に比べてより一般的であろう．また，マウスをサポートしているものもこの中に含む．2つ目として，WWWブラウザに対応して，英語の原文に単語の意味を直接埋め込むタイプがある(埋め込み型)[3]．3つ目は原文をそのまま残して，違うページまたは原文に隣接させて日本語の意味を記すタイプがあるだろう(ページ分け型)．

我々もまたリファレンスシステムを開発し，そのシステムをAutorefと名付けた．上記した分類に従うと，Autorefはリファレンス型，Proxy型，ページ分け型に属する．

我々はこのようにリファレンスシステムの分類を行なった．ここで，これを現時点で開発されているシステムに割り当ててみよう[4][5]．これを表1に示す．

第3節では我々が開発したシステムAutorefの概要を説明する．そして第4節では，本節で述べたリファレンスシステムのトレードオフをその分類により比較し説明したい．

3 システムの概要

Autorefを開発するにあたり，我々はProxyサーバ[6]の機構に注目した．Proxyサーバは元来セキュリティのために開発されたもので，外部とIP接続が制限されているマシンのWWWブラウザの代理をして，外

部接続を行なうシステムである．この時，接続の代理を行なうだけでなく，日本語のコード変換などの付加をするサーバ (delegate)[7] もある．ローカルホストにProxy サーバを設定した場合，WWW ブラウザではなく，そのProxy サーバがリモートホストのWWW サーバと見かけ上交信することになる．よってブラウザはProxy サーバとだけ交信して情報を得る．そして，WWW ブラウザとProxy サーバの間に付加的なシステムを置くことも可能である．本システムはこの位置に置かれる．

WWW では多種の情報を扱うことが可能であるが，Autoref は HTML(Hypertext Makeup Language) ドキュメントのみを処理する．システムは HTML ドキュメントのフォーマットを知っているから，Proxy とブラウザの情報の交信を仲立ちする時に，それ以外の情報はそのまま通過させ，HTML ドキュメントだけを処理する．

ユーザの要請により，WWW サーバから HTML ドキュメントが送られてくると，システムはそのドキュメントの中でリファレンスを付加するべき単語を検出して，そのURL(Universal Resource Locator) リンクを付加する．その結果，WWW の本来あったリンクに加えて，Autoref が付加したリンクをユーザは利用できる．これにより，ユーザは付加したリンクをクリックするだけで，電子英和辞書のように単語の意味の情報を得ることができる．また辞書を換えることで，様々なリファレンス情報を見ることができる．図1にリファレンスシステム Autoref の概観を示す．

ソフトウェアは3つのモジュールから成立している．Proxy サーバを通過した情報が HTML であることを識別するモジュール，HTML の構造に従ってリファレンスの追加が可能な部分を取り出すモジュール，リファレンス情報を付加する単語を選択するモジュールである．システムは単語とリファレンス情報を対応させるURL のペアを持っている．各単語は最長一致で対応を取り，該当する単語を見つけた場合には，その説明のドキュメントのありかである URL を求める．そしてこの URL をそのまま HTML に挿入する．オリジナルの HTML ドキュメントとの違いは対応する URL があるかないかである．

Autoref の英和辞典は，通常の辞書とは異なり，単語の変化形についても独立の見出し語を生成してある．それは，原文から辞書の登録単語の有無を判定するから，変化形についても見出し語が必要なためである．変化形を正しく求めるのは，一般には簡単なことではないが，規則に従って変化形を作成した．さらにこれをスペルチェッカに通し，正しい語を選択するようにしている．

Autoref はマルチユーザ環境を提供し，そしてドキュメント中の全ての単語を処理する．それゆえ，単語を辞書引きする速度は重要な問題である．従って辞書引きの効率を考慮してシステムを構築した．そこで，全ての登録単語をメモリ上に展開させ，ディスク操作を避けた．また単語情報は木構造を使ってメモリ上に格納するようにした．Autoref は単語を1文字づつ判定しノードを状態遷移する．この構造はメモリ空間を消費するかもしれないが，最長一致の語を取るのに適しており，登録単語を速く辞書引きできる．また Autoref はカスケード接続により，数種の辞書を混合して持つことが可能である．

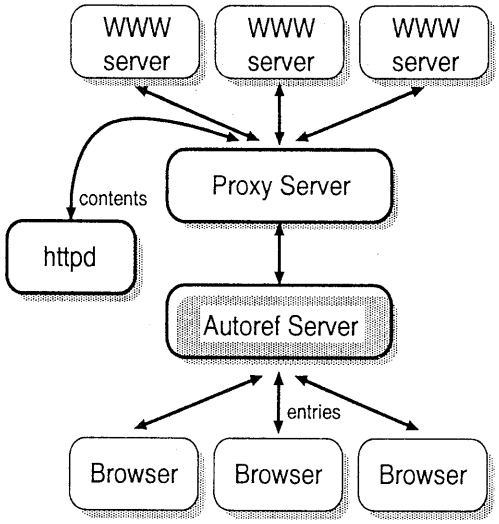

図 1: Autoref の概観

4 ユーザの見地からの比較

リファレンスシステムの優劣を決定するのはユーザである．各システムにおいてこの事に変わりはないが，その評価は一般に難しい．それはシステムの数，種類が多く，その比較項目も数あるためであろう．最近のシステムは専門辞書を数種，多いものでは十数種持っている．辞書は，専門辞書の種類，数，語数，整理の度合などがシステム毎に異なる．また一度に扱える辞書の種類もシステムによって異なる．さらにシステムがマルチユーザに対応するか，自動翻訳の場合はその翻訳精度など，比較項目は多い．そのため，リファレンスシステムの優劣を決定するのは無理があるように感じる．しかしながら，ユーザがドキュメントを読む時間とその辞書引き回数は重要な指標になると我々は考えている．そこで，第4節と第5節では第

2節で示した分類に従い，辞書引き方法を比較し，今後のシステムの方向性を考えてみたいと思う．まず，第4.1節でWWWへの関与の形として分類した汎用型とブラウザ型とProxy型の3つを比較し，次に第4.2節でリファレンス方法として分類した自動翻訳型とリファレンス型を比較してみる．

4.1 WWWへの関与の形

最近ではパーソナルコンピュータ上に多くの辞書引きシステムが存在する．汎用型はパーソナルコンピュータ上でも，ネットワークにつながっているワークステーション上でも問題なく利用できることが強みである．しかもWWW上のドキュメントのみならず，ドキュメント形式のものは全て扱うことができる．しかしながら，WWWへの対応を考えた場合，ブラウザ型やProxy型などの専用的なものに比べその操作性はやや劣るだろう．

WWWブラウザにリンクしてリファレンス機能を強化したソフトウェアが市販されている．これは我々がブラウザ型と分類したものである．またこの事はリファレンス情報がWWWでの形態で有用であると解釈することもできる．具体的には，英和辞書システムと自動翻訳システムとを組合わせたものである．ブラウザ型の場合，WWWブラウザと密接にリンクしているため，WWW専用の機能が組み入れられているのが特徴である．例えば，自動翻訳ソフトの中には，ブラウザがディスクにキャッシュしたデータを読み込めるものがある．従って，ユーザはサーバにアクセスしてキャッシュにデータを貯めてから回線を切断し，その後で翻訳するという使い方ができる．

ここではWWWの使用を前提においてブラウザ型とProxy型を比較してみる．Proxy型は汎用のブラウザをそのまま利用するようなシステム形態を採用している．汎用ブラウザをそのまま利用できる利点はいくつか存在する．まず，WWW自身の変化への追従が容易なことがある．使用する情報の種類はWWWにおいて今なお増加している．動画を扱う拡張もなされている．このような背景からブラウザは進歩の途中であり，ブラウザ型はブラウザの多様な変化に追従する必要があるため，今後の開発コストも大きいだろう．リファレンスシステムを構築する際，ブラウザには多くの種類とバージョンがあるため，全てのブラウザに対応させることは現実的でない．たとえ全てのブラウザに対応したシステムが作られたとしても，それらを更新していくことはさらに難しいし，システムの入れ換えが面倒になる可能性がある．しかし，Proxy型は既存のブラウザから独立しているから，そのような煩わしさはない．このような背景で，ブラウザは汎用品のままでリファレンスの機構を実現するのは，実際の問題として有用である．このような観点から，Proxy型はネットワーク接続が可能なコンピュータ上では汎用的に有利であると言えるだろう．現在NetScape Navigator^{TM}(米国Netscape Communications Corporation)が世界標準のWWWブラウザになりつつある．ブラウザ型は実際には，このNetScapeの仕様に合わせたものがほとんどである．

4.2 リファレンス方法

コンピュータの高性能化とWWWの普及に伴い，現在，多くの自動翻訳ソフトが市販されている．最近では低価格の翻訳ソフトが発売されているため，個人でも十分手の届く範囲にある．いろいろなタイプのものが発売されている．中にはAutorefと同じように，Proxyを設定し全てのWWWブラウザに対応するものもある[8]．これらのソフトウェアは，翻訳精度を上げると翻訳速度が犠牲になり，翻訳速度を上げると翻訳精度が犠牲になる．文献[4][5]を参考に英文の翻訳速度を計算して見た所，Intel社のPentiumをCPUとして用いた場合に，翻訳精度の荒いもので約50語/秒，翻訳精度の高いものでA4紙1ページあたり約1分で翻訳結果が得られるようである．これらのシステムのほとんどは基本辞書と専門辞書などを組合せて，数種の辞書を一緒に参照することが可能になっている．

WWWへ関連したProxy型の開発がなされないうちは，リファレンス型は，ほとんどが単語をタイプすることによってその意味を得る汎用的な辞書引きシステムでしかなかった．これを我々はリタイプ型と称したが，ブラウザ型の自動翻訳ソフトは，この辞書引きシステムも同時に実現して辞書引き機能を補っているものが多い．

リタイプ型の場合，特定の単語の意味はその単語をタイプすることで得られる．このタイピング操作は，しかしながら，多くのドキュメントを読む時には時々面倒になる．また，WWWを利用する場合などはブラウザの基本操作であるマウスクリックに適合せず不便である．ユーザがドキュメント中の単語を指定するのにマウスをサポートしているものもある．これによりユーザは単語をタイプしなくてもその意味を得ることができよう．しかしながら，ユーザは辞書がその単語を含まないために，辞書引きを失敗するかもしれない．WWWにリンクし，Proxyを利用したリファレンスシステムの新進により，このリファレンス型が新しく変わってきている．この型は，辞書にない単語は訳されずに表示されるだけなので，少なくともリタイプ型のようなストレスは感じないだろう．

リファレンス型と翻訳型の比較はユーザの使用目的により評価が変わってくるため，厳密な比較は難しい．そこで我々は日常的な英語の話題を理解するという基準を設けて，辞書引き方法を比較してみることにした．また我々は今後のシステムはWWWへの関連を密にすると考え，ブラウジング方法はProxy型のページ分け型と埋め込み型の両方に対して比較を行なうことにした．これにリタイプ型と翻訳型を合わせてその4つを比較する．この詳細を第5節にて説明する．

5 辞書引き方法の比較検討

我々はリファレンス型のブラウジング方法であるページ分け型，埋め込み型，リタイプ型の3つに翻訳型を加えた4つのリファレンス方法の英文辞書引き実験を行なった．

辞書はフリーソフトウェアのEDICT[9]を使用した．これは一般的な辞書で単語数は18,144である．ページ分け型には，我々が開発したAutorefを使用する．埋め込み型には，ページ分け型のAutorefと辞書を共通にするために，システムを変更し，Autorefの埋め込み型を作成して使用することにした．リタイプ型も辞書を同じにして我々が作成した．翻訳型にはWWWブラウザ対応の自動翻訳ソフトNetSurfer/ejを使用した．これは市販されている製品であり，使用許諾条項により，その辞書を改変することはできないため，このシステム附属の単語数約70,000の一般的な辞書を使用した．従って，翻訳型とリファレンス型の実験結果を直接比較することはできない．

日常的な英語という基準から，WWW上に掲載されたCNNのニュース記事[10]を実験用ドキュメントとして使用した．米国内のリアルタイムな話題を使用したため，被験者にとってあまり馴染みのない話題である．また4人の被験者は全て日本人の大学生であり，その英語力もそれ相応である．

WWWブラウザを使用して，4つの記事を4つのリファレンス方法で被験者を換え英文を訳すことにした．つまり被験者は各英文記事を違うリファレンス方法で訳すことになる．リファレンス方法と記事と被験者の関係を示したものを表2に示す．また記事のURLは付録Aに示してある．

実験で計測したのは，その記事を訳すのにかかった時間，および，システムから十分に情報が得られず，他の辞書を使用した回数である．つまりこの事は，単語の意味がシステム辞書にない場合や，英文または単語が適切な和訳をされていない状況を示す．また他の辞書を利用する場合には，各自利用する英和辞書を全ての記事で統一してもらった．そして，この実験で「記事を訳すのにかかった時間」とは，被験者が記事にアクセスし，その内容を理解し納得して読み終えるまでの時間である．その結果を表3に示す．

表2: 実験の関係表

リファレンス方法		英文記事 1^1	2^2	3^3	4^4
翻訳		A	B	C	D
リファレンス	ページ分け	B	C	D	A
	埋め込み	C	D	A	B
	リタイプ	D	A	B	C

A～D ・・・被験者
記事の語数 ・・・ ^{1}502 ^{2}481 ^{3}697 ^{4}483

表3: 実験結果

リファレンス方法		英文記事 1	2	3	4	計
翻訳		55†	14	29	21	119分†
		36‡	5	15	10	66回‡
リファレンス	ページ分け	30	31	34	37	132分
		15	10	4	17	46回
	埋め込み	31	25	45	22	123分
		30	8	34	18	90回
	リタイプ	14	46	57	34	151分
		4	16	17	10	47回

† 上段 ・・・記事を訳すのにかかった時間[分]
‡ 下段 ・・・システムの辞書以外で単語を調べた回数[回]

表3の数字だけを考慮すると，Proxyを用いたページ分け型と埋め込み型がリタイプ型よりは良いような印象を受ける．それは手の動作量を考えればうなずける．では，翻訳型とリファレンス型はどうだろう．翻訳型の辞書は約70,000語の語彙を持っており，これはリファレンス型の辞書の約3.8倍である．よって直接比較することはできないことは前に述べた．ここでは数字以外のことも考慮して比較する．

翻訳型は英語のドキュメントを和訳することにより，英文を読むきっかけをユーザに与えている．よって，英文を斜め読みするということではかなり適していると言える．しかしその反面，誤訳の多さにより合計66回も辞書引きをしなくてはならなかったという事実もある．英語を和訳して理解するという前提での誤訳は，英文だけを示されることよりもユーザにストレスを感じさせる．従って，英語の能力が高いものならば誤訳の多い自動翻訳に頼るのは煩わしいだろうし，辞書引き回数も減るから，リファレンス型の方が時間の節約につながるだろう．ユーザが翻訳ソフトによる英文の理解を求めるならば，今後のさらなるコン

ピュータ性能の向上と翻訳精度の向上が必要であるだろう．これらのことから，翻訳型とリファレンス型のどちらかが優れているとは言えないと我々は考えている．

しかしながら，WWW との関連を考えた場合，我々は Proxy 型を強く推薦する．これを生かすには使用辞書の改善が必要であるが，HTML ドキュメントとリンクすることにより，気軽に単語の意味を調べれるのが強みである．さらに，言語の辞書だけでなく，一般的なリファレンスシステムを構築しようと考えた場合，Proxy 型のシステムはリファレンスの入れ換えに容易に対応するから，その拡張性は高い．

我々は辞書引き方法を評価する環境を作り，その一部を評価したが，満足する客観的データは得られていない．よって，この結果だけで辞書引き方法の優劣を断定するのは危険であるが，その方向性は見えたような気がする．今後さらに実験を重ね，その不足を補いたい．

6 まとめ

我々は多種存在するリファレンスシステムを WWW への関与の形，リファレンス方法，ブラウジング方法で分類した．また我々は Proxy を利用した WWW 上でのリファレンスシステム Autoref を開発したことを述べた．本システムを使用することにより，ユーザはマウスクリック一つで単語情報を参照することができる．その特筆すべき特徴は，ブラウザの種類を問わずに使うことができる点にある．

それから我々は分類に基づき辞書引き方法を比較した．具体的には，人が WWW ドキュメントを理解する時間と辞書引きした回数を計測した．この実験だけで辞書引き方法の優劣を断定することはできないが，定量的なデータを考慮し，今後のシステムの方向性を示した．客観的データを確立するためのさらなる実験と，本システムが英和辞書以外のリファレンスシステムを構築することが我々の次の仕事である．

謝辞

本研究は NTT の工学研究育成の援助を受けています．また本研究の実験に際して手伝って下さった本研究室の松本兼一氏と増田恵子氏および本学 VLSI 設計研究室の小松平良樹氏に感謝します．

参考文献

[1] 吉村伸．インターネットの利用と仕組み．*UNIX MAGAZINE*, pp. 49–55, 12 1995.

[2] T. Berners-Lee, R. Callian, A. Luotonen, H. F. Nielsen, and A. Secret. The World-Wide Web. *Communications of the ACM*, Vol. 37, No. 8, pp. 76–82, 1994.

[3] 鵜川義弘，秋元学，藤田信之，佐藤豊．DeleGate Proxy を使った Web 情報の英和逐語訳サーバ -EtoJ_Proxy-. In *Japan World Wide Web Conference '95*, 11 1995.

[4] 広野忠敏．最新・英文翻訳ソフト．*INTERNET magazine*, pp. 120–127, 12 1995.

[5] 渡邊昭雄．パソコン翻訳の時代．*bit*, pp. 98–104, 9 1995.

[6] Ari Luotonen and Kevin Altis. World-Wide Web Proxies. In *WWW'94 Conference*, April 1994. <http://www.w3.org/pub/WWW/Proxies>.

[7] Yutaka Sato. Development of a Protocol Mediation System Delegate. Technical Report in ETL TR-94-17, Densouken kenkyusokuhou, 1994. (in Japanese), <ftp://etlport.etl.go.jp/pub/DeleGate/ETL-TR94-17.ps.gz>.

[8] Toshiki Murata, Hideki Yamamoto, and Junji Nagata. WWW machine translation system: W3-PENSEE. *SIGNL Research Report 108-22*, pp. 149–152, 1995. (in Japanese).

[9] ftp://ftp.cc.monash.edu.au/pub/nihongo.

[10] http://www.cnn.com/index.html.

A 実験で使用した英文記事の URL

記事 1 Harlem massacre may have been race-related
http://www.cnn.com/
US/9512/harlem_fire/12-09/index.html

記事 2 Homeless shelters may face funding freeze
http://www.cnn.com/
US/9512/shelter/index.html

記事 3 GOP and White House dig in for another budget standoff
http://www.cnn.com/
US/9512/budget/12-08/pm/index.html

記事 4 Speed limits on the way up
http://www.cnn.com/
US/9512/speed_limit/index.html