ヒューマンインタフェース 75－6
（1997. 11. 14）

ネットワークを利用した辞書サービスシステム

木村 和広　平川 秀樹

東芝 研究開発センター

情報・通信システム研究所

{kim,hirakawa}@eel.rdc.toshiba.co.jp

概要

日本語IMEなどの言語処理アプリケーションが利用する機械処理用辞書に対して、辞書更新情報をネットワーク経由で配信する、辞書サービスシステムについて述べる。既成の辞書情報は、固定化されたものであり、最新用語や組織内用語などは含まない。これら多くの未登録語の存在は、実文書を処理する上で、処理品質を下げ、ユーザが期待を裏切る結果をもたらす。そこで、これらの用語を収集し、アプリケーションにオンライン自動登録する枠組を提案する。用語は提供側が用意するだけでなく、ユーザからも獲得してこれを活用する。本サービスにより、これまで個人レベルで行なっていた辞書カスタマイズ作業の負荷が軽減され、アプリケーションの品質維持に貢献する。

A Dictionary Service System on Computer Networks

KIMURA Kazuhiro　HIRAKAWA Hideki

Communications and Information Systems Laboratories,
Research and Development Center, TOSHIBA

Abstract

This paper describes a dictionary service system on computer networks. It provides updated dictionary data for natural language processing system, such as Japanse input method editors(IMEs). Because the system dictinary of the target system does not provide up-to-dated words or commonly used words in a local community, the target system does not always produce a satisfying result when it processes real documents. The system proposed here also gives a framework for dictionary sharing among users. As a result, the system reduces individual effort for dictionary customization and maintains the system performance.

1 はじめに

計算機への日本語の入力を行うIMEや、文書を翻訳する機械翻訳システムなどが広く利用されるようになってきている。この種のソフトウェアは、種々の言語情報を含む辞書を利用している。しかしながら、既成の辞書情報は、固定化されたものであり、次々に生まれる最新用語などに追随することができず、また、会社組織などローカルなコミュニティで使用される用語などは含まない。

このため、ユーザは、個人辞書に、所望の単語を登録することにより対応してきた。しかし、このようなカスタマイズ作業は、心理的・物理的負荷が多いため、不便さを我慢しながらも、カスタマイズなしにシステムを利用している一般ユーザも多い。

本稿で述べる辞書サービスとは、計算機ネットワークを利用して、ユーザの所望する用語を、ユーザの持つアプリケーションにオンライン自動登録するサービスである。あるいは、個人の辞書カスタマイズを代行支援するサービスと言ってもよい。

2 辞書知識の共有とサービス

辞書知識の開発は、コストが高い。これは、人間の介入が避けられないためである。そのため、辞書サービスを実現するためには、辞書知識の効率的獲得技術の深耕が重要であるが、それと並んで、得られた知識の共有が重要である ([1],[2])。

図1は、辞書知識共有の枠組を示したもので、大きく分けて、3つのサブシステムから構成される。収集系・管理系が、辞書サービスの背景にある要素技術である。

収集系サブシステム 時代に即応した用語の収集・提供、広範な分野の専門的語彙や特定集団内での用語の獲得を行うために、電子化テキストからの半自動的な新語抽出を行なう。特に、WWWホームページやネットニュースなどには、時事的な単語の出現が多いため、インターネットロボットなどと連携して、定常的な新語抽出・蓄積を実現する。

管理系サブシステム 収集系サブシステムにより獲得された言語情報を種々のアプリケーション間で共同利用できる形態で、管理する。このため、応用には依存せず、流通に適した構造をもったデータベースが必要で、Text Encoding Initiative(TEI[3])の推奨フォーマットに準拠した構造を採用している。

分配系サブシステム (=辞書サービスシステム) 管理系サブシステムに蓄積された言語情報を、インター/イントラネットを経由して、エンドユーザの持つアプリケーションシステムに、オンライン自動登録する仕組みを提供する。辞書知識の知識源としては、提供側がトップダウンに用意したものだけでなく、ユーザが個人で蓄積している辞書データの活用も考えられる。そこで、ユーザの個人辞書情報を自動アップロードし、分野別辞書、組織内用語辞書などの形態に集約し、再分配する。

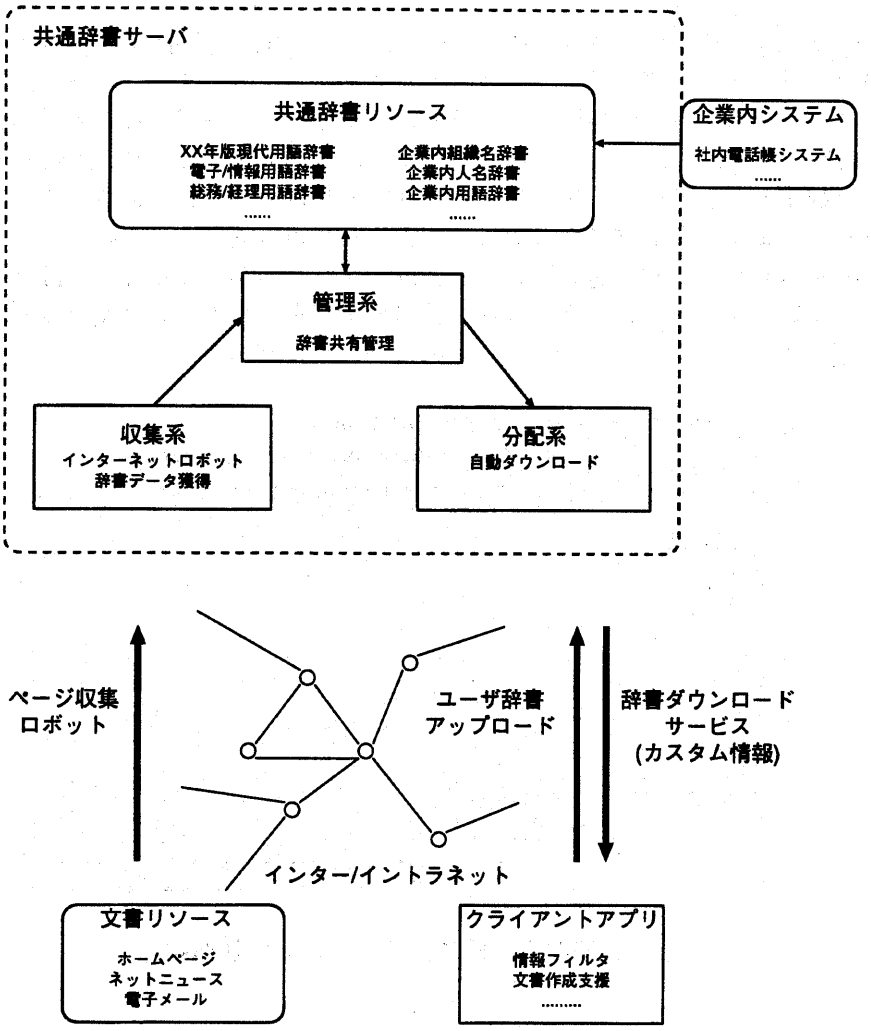

図 1: 辞書知識の共有

以下、辞書サービスシステムに焦点をあてて述べる。

3 辞書サービスシステム

辞書サービスシステムは、クライアントーサーバ型のシステムである。クライアントシステムは、辞書更新処理がユーザが意識することなく行われるよう、インタフェース設計されている。すなわち、一度初期設定を済ませれば、特に操作を必要としない。

3.1 クライアントシステムのGUI

図2は、初期設定のためのGUIである。日本語IME、翻訳システムといった辞書更新のターゲットアプリケーション別に設定を行なう。まず、ユーザは「辞書選択」に示される種々の辞書セットのうち、所望のものを選択する。各種辞書は「専門語–情報–ハードウェア」などのように階層的に構成されており、上位階層のボタンを選択すると、その下位に位置するすべての辞書セットが選択される。なお、辞書セットの種類や階層構造の情報は、辞書サーバから提供され、動的にメニュー化されて表示される。辞書選択メニューには、ユーザの要望等による、辞書セットの追加や変更に対応できるよう、動的な追随性が要求される。その他の設定項目として、更新タイミングや、アップロードの可否の設定がある。

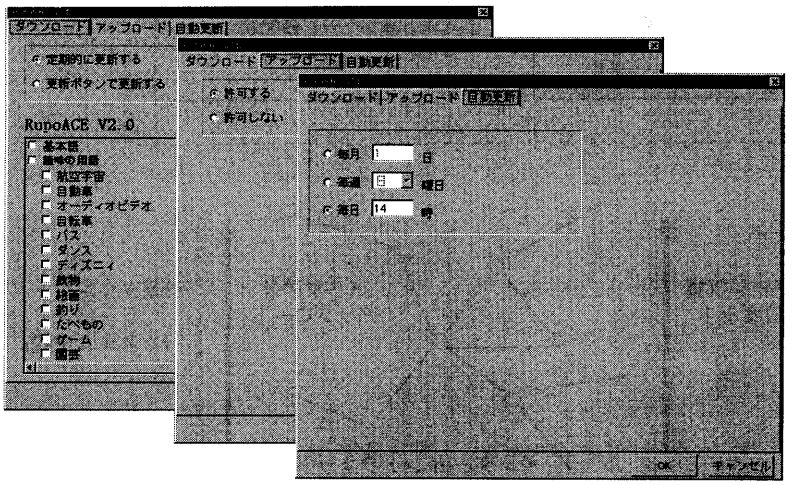

図 2: 辞書サービスクライアントの初期設定

3.2 辞書サービスシステムの動作

クライアントは、指定された日時に起動され、サーバとの接続を行ない、最新の辞書セット情報を獲得する。そして、クライアント側に保持されているダウンロードログを参照して、どの辞書セットから何語登録されているかを調べ、辞書サーバに通知する。同時に、ユーザの設定情報を調べ、ダウンロード要求する辞書セット名を通知する。辞書サーバは、これらの情報から、要求された辞書データから、既にダウンロードされている辞書データを除いた、差分の辞書データを決定して、これを送信する。続いて、ターゲットアプリケーションに対して、受信した辞書更新情報を与え、そのシステムの辞書を更新する。辞書更新情報は、追加項目と削除項目からなり、用語の追加だけでなく、死語の削除も可能である。

さらに、辞書アップロードが許可されている場合、ターゲットアプリケーションに対し、ユーザが登録した新しい単語のリストを要求し、これを獲得する。この場

合、前回獲得した分との差分が獲得されることが望ましいが、通常、既存のターゲットアプリケーションは、単語の登録日時など、差分を判定するための情報を保持していることはまれである。そこで、その時点のユーザ登録語の全リストを獲得し、その獲得日時と共に保持、前回までの全単語リストと、今回の全単語リストを比較し、その期間に新たに登録された単語のリスト（差分）を作成する。もし、差分がある場合は、辞書サーバに対して、その差分情報を送信し、セッションを終了する。サーバは、クライアントから受信した辞書データを一旦プールしておく。そして、一定時間ごとに、プールされた全データを集めて、頻度集計し、ある閾値を越えるデータを新たな再分配単語候補として選定する (多数決)。これらの候補は、辞書編集者のチェックを経て、ダウンロード用データに追加される。なお、ユーザから獲得した辞書データには、ユーザがダウンロード設定した分野情報が付加されており、獲得語彙の分野推定に寄与する。

3.3 階層型辞書サーバ

辞書サーバは、図3のように階層構成することもできる。階層型のサーバ構成は、特に大きな組織内で辞書サービスを運用する際に適用することにより、組織内の用語シェアの効率化やサーバの負荷分散を図ることができる。

図に示すように、辞書サーバは、会社組織などの各部門に対応して、部門サーバ、上位部門サーバ、組織内マスターサーバと、階層的に構成される。ユーザは、リーフサーバをホストとして利用する。一般の分野別辞書については、マスターサーバの保持するダウンロードデータのコピーが、一定時間ごとにリーフサーバまで伝達され、ユーザに供給される。一方、ユーザからアップロードされたデータは、各部門サーバごとに、多数決により辞書化され、部門共通辞書となる。部門共通辞書は、そのままユーザに提供される他、上位部門サーバに送信される。上位部門サーバは、各部門サーバから収集した部門共通辞書を、やはり多数決によって辞書化し、上位部門共通辞書を作成する。このような構成は、組織内用語の自動分別を可能にし、ユーザは、自部門で共通的に使用される単語だけでなく、組織内で共通的に使用される単語もまた、自動的に利用できるようになる。

4 まとめ

本稿では、計算機ネットワークを利用した辞書のサービスについて述べた。本サービスは、近々インターネット上で公開し、ユーザからのフィードバックを受ける予定である。今後は、サービス対象とする辞書知識を用例知識、用語の定義文 (人間が参照するための辞書) などに拡張していく予定である。また、クライアントシステムは、サーバとの交信だけでなく、自らテキストからの用語獲得機能を包含するなどして、ユーザに届くメールや、ユーザが参照したWWWページをリソースとした個人辞書カスタマイズを実現し、自律的な辞書エージェントとして、昇華させてゆきたい。

イントラ向け分配収集サーバ

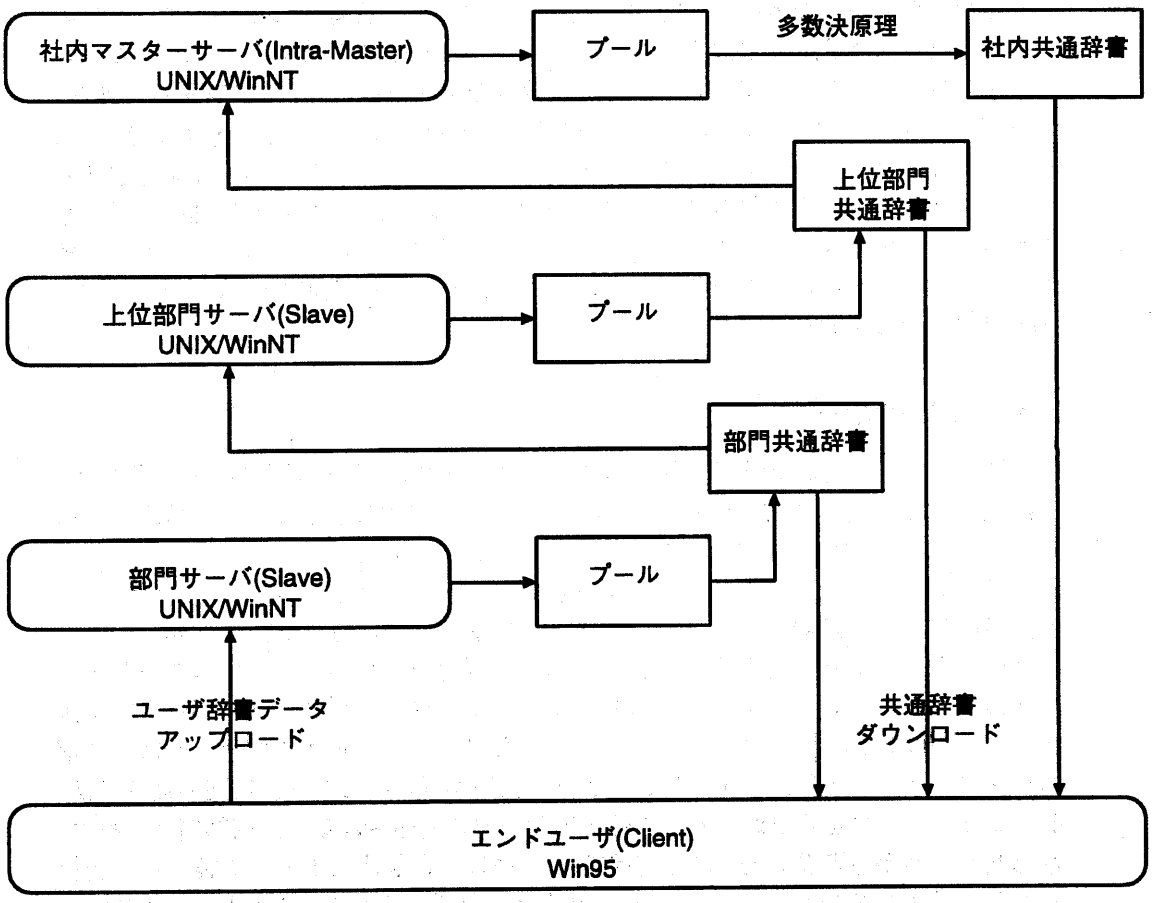

図 3: 階層型辞書サーバ

参考文献

[1] 機械翻訳ユーザ辞書の共通フォーマットの設定 言語処理学会第 3 回年次大会論文集, pp.19–22 (1997).

[2] 淵 武志: WWW 上での言語データ収集環境, 言語処理学会第 3 回年次大会論文集, pp.139–142 (1997).

[3] Sperberg-McQueen, C.M. and Burnard, L.: Guidelines for Electronic Text Encoding and Interchange(TEI P3), (1994).