ネットワーク上の専門用語辞書の統合化

木戸 冬子 (マイクロソフト アジア リミテッド, E-Mail: fkido@microsoft.com)
中川 裕志 (東京大学情報基盤センター, E-Mail: nakagawa@r.dl.itc.u-tokyo.ac.jp)

1 はじめに

電子化辞書としては，従来より新聞記事などの一般的文書における語彙を対象にした大型の辞書(EDR[1], NTT[2], IPAL[3])が開発され，研究などに使用されている．しかし，ビジネス活動は，その目的からして当然のことながら専門性が高い．そのようなビジネス活動の場においては，多くの場合，専門用語がコミュニケーションの媒体として重要な役割を担っている．パソコンの注文における仕様にしてもOS名(Windows 2000)，周辺機器名(DVD-RAMドライブ)とその特性 (IEEE1394インターフェース)，機器の形態 (Desktop)など全ての情報が専門用語によって記述される．したがって，整備された専門用語の電子化辞書は，ビジネスの効率化のための基本的情報リソースである．残念なことにひとつの組織においてさえ，このような専門用語辞書は，一定の形式で統一された情報リソースとして整備されていないことが多い．一方で，小さな辞書が分散的に開発されている．その結果，情報としては存在するのに，同じ組織内でもアクセスができないという事態が発生する．このような問題点は，自然言語処理の理論によってすぐに解決するわけもなく，まず基本的な専門用語を網羅する統一された形式の辞書を，人手で開発しなくてはならない．このとき気をつけなければならないのは，統一された辞書ができるだけ多くの目的に適用できる多機能辞書にすることである．本研究では，マイクロソフト社内のネットワーク上に散在しているさまざまな専門用語辞書を統合し，多機能辞書を構築したので，上記の問題に対する解決策の一例として報告する．

2 インターネット上の辞書の現状

2.1 Web辞書の機能

現在，インターネット上には多数の辞書が氾濫している．インターネット上の辞書 (以下，Web 辞書と記す) は，書籍の辞書と同様，英和・和英などの訳を検索する辞書と用語の意味を検索する用語辞書に大別される．Web辞書はインターネットという性格上，コンピュータ関連 (パソコン，通信，情報，マルチメディアなど) の専門用語に関するものが多いのが特色として挙げられる．Web辞書を含む電子化された辞書と書籍の辞書とを比較した場合の違いは，キーワード検索，マルチリンガル検索，クロス検索などの多様な検索機能と，テキストデータ以外に静止画，動画，音声などのデータが扱えることである．

これらの Web 辞書独自の機能は，書籍の辞書よりも，電子化された辞書を選択する理由ともなっている．

2.2 Web辞書の見出し語数

電子化辞書を評価する指標として，網羅性，汎用性，メンテナンスビリティがある．また，用語辞書としての評価については意味記述の品質も重要である．

Web 辞書を網羅性の評価するため見出し語数と見出し語の重複率を評価した．対象としたのは，コンピュータ関連の Web 辞書として代表的なアスキーデジタル用語辞典[4]，通信・情報辞典 e-Words [5]，パソコン用語知ったか辞典[6]の３つである．調査の結果，見出し語数が最大のアスキーデジタル用語辞典と最少のパソコン用語知ったか辞典とでは，倍以上の差があった(表 1)．更に，アスキーデジタル用語辞典に対して，重複率 (他の辞書の見出し語がいくつ含まれているか)を調査した結果，通信・情報辞典 e-Words が 24%，パソコン用語知ったか辞典が23%となった．これから，見出し語数の多いアスキーデジタル用語辞典の網羅性が，単純に良いとは言えないことがわかる．重複率がどちらも 20%台となった理由は，異表記問題 (”109 日本語キーボード”と”109 キーボード”など) があった．

また，調査対象となった辞書すべてにおいて，2000 年 8 月に調査した値よりも，見出し語数が増加していた．このことは，Web 辞書の新語の多さと更新頻度の高さを表すとともに，メンテナンスビリティの重要性を再認識させる結果となった．

2.3 Web辞書の内容

Web 辞書は，記述項目にばらつきが多いのも特徴である．調査対象の辞書において共通なものは，見出し語，意味，別名であった(表 1)．

辞書を統合するという観点で考えた場合，最も処理が難しいのが意味である． 例えば，表 1 の記述内容を例にとった場合，意味以外の項目は，最大公約数をとる事で対応できるが，意味は同様の方法では対処できない．また,記述内容の品質も考慮しなければならないため，機械的に処理するのは困難である．

表 1 Web 辞書と記述内容（2001 年 2 月 9 日現在）

辞書名	項目	見出し語数	重複率
アスキーデジタル用語辞典	見出し語，別名，意味，関連語	5,354 (3,083)	—
通信･情報辞典 e-Words	見出し語，正式名称，別名，読み方，意味，関連語，関連資料，カテゴリー	3,798 (2,738)	24% *[893]
パソコン知ったか辞典	見出し語，意味，別名	1,775 (1,543)	23% *[411]

()内は 2000 年 8 月のデータ

* []内は，アスキーデジタル用語辞典と重複した見出し語数

3 多機能辞書の構築

3.1 多機能辞書に期待される役割と機能

ビジネスモデルとして社内利用を想定した場合，専門用語辞書には以下の役割が期待される．

1）用語辞書機能

テキストデータを検索するためのキーワードとして専門用語を使う場合，辞書によって正確なキーワードを得ることは必須である．これは網羅性の良い用語辞書機能を要請する．

2）シソーラス機能

同じ意味内容を別の用語で表現することはしばしば起こる．ところが，このような状況は異なる部署で異なる用語を使うという現象として発生するため，ややもすると部署間での意思の疎通を欠く原因となりやすい．そこで，同義語や関連語，さらには縮約表現を含めたシソーラス機能が必要となる．

3）カテゴリー機能

データベース検索においてカテゴリー検索も多用されるが，この場合には見出し語のみでなく，見出し語をいくつかのカテゴリーに分類してあることが望ましい．また， 洗練されたカテゴリーは，問題の自己解決を目的とするヘルプシステムなどの将来的な課題に適応するためにも重要である．

4）縮約された表記

縮約された表記は同義語の一種であるので，用語辞書機能としても重要だが，携帯端末でのテキスト表示などの課題にも適応するためには，縮約表記を同義語として登録しておくことが重要である．特に最も短い表記をマークしておくことは高速アクセスの観点からも必要なことである．

5）日英対訳

現在，多くの情報が英語で提供されており，日本語で提供されていない情報も少なくない．適切な対訳キーワードを用い，日本語のキーワードで英語の情報を検索できるようにすることは，情報検索という観点において重要なことである．

3.2 辞書リソースの統合の手順

辞書の統合に関する先行研究としては，羽田，松本らの自然言語処理用の汎用辞書構築がある[7]．羽田，松本らの研究は，語数の不足と情報の欠落を補完するという観点で類似性があるが，自然言語処理の汎用辞書を目的としているため，本研究の目的とは異なる．

本研究における辞書リソースの統合は，

1) 統合可能な辞書の選択

2) 見出し語の統合

3) 同義語の関連付け

という手順で行った，まず，社内に分散している辞書のうち統合できるものを選んだ．辞書リソースとしては，製品のヘルプなどに付属している用語集の他，社内ネットワーク上の言語リソースとしておのおの多数の辞書からなるGlossary, New Term List, Terminologyと呼ばれる 3 群の辞書がある．今回は，これらの辞書群を統合して，新たに多機能辞書とした．

見出し語の統合は，日英を区別せず最大公約数をとる形で行い，個々の見出し語に言語属性を設定した．また，正式名称，対訳，頭文字語などはすべて見出し語として登録し，見出し語同士の関係を，同義語，頭文字語，上位語，下位語，類義語，関連語というレベルで関連付けた．また，言語属性の設定および見出し語の関係付け作業は，情報の選別(screening)をかねて，Windows Access 2000で作成した辞書メンテナンスツールを用い，人手で行った(図1)．また，意味データの優先順位は，製品，社内の言語リソースの順とし，更に更新日時が新しいものを優先して用いた．

3.3 用語辞書の公開

多機能辞書の一部を新Terminologyとして社内公開した(図2). 新Terminologyは，意味の検索，関連語のクロス検索の他，見出し語から技術情報をダイレクトに検索できるサービスも提供している．また，検索した単語が見つからなかった場合には，その単語の登録リクエストができる機能も提供した．

新Terminologyの見出し語数および項目は，使用した言語リソースと比較する形で表2に記載した．

表2 言語リソースの一例と新 Terminology の比較

言語リソース名	項目	見出し語数
Windows 2000 ヘルプ (用語集)	見出し語，異表記，意味，関連語	805
Glossary (Active Directory)	見出し語，異表記，意味，関連語	69
New Term List	見出し語(英語)，対訳(日本語)，プラットフォーム，製品	67
Terminology	見出し語，正式名称，意味	410
新 Terminology	見出し語，フリガナ，頭文字語，登録商標，固有名詞，言語，同義語，上位語，下位語，関連語，読み方，意味，縮約表記，カテゴリー	6,190

4 おわりに

本稿では，社内ネットワーク上に散在している専門用語辞書を統合し，多機能辞書構築の試みについて述べた．

今後の課題として，以下を検討している．

1) 粒度問題の解消

意味記述において，粒度問題がある．例えば，句読点（", ．", "、。"）などの比較的単純な異表記の問題から，文体（"です･ます調", "だ･である調"）など記述内容の詳細さの違いや使用する単語の違いなど，見出し語の出典によりさまざまである．異表記は機械的な対応も可能であるが，それ以外は機械的な統一化は困難である．今後，辞書公開する場合に，本問題の解消は不可欠である．

2) 利用者層の拡大

現在登録されている見出し語は社内の言語リソースを使用していることから，ある意味で専門技術者のための専門用語である．しかしながら，今後，更なるビジネス的活用（顧客対応）に活用することを考えた場合，コンピュータの専門技術者ではない層も取り込まなくてはならない．例えば，ACPI (Advanced Configuration and Power Interface)という用語を知らない場合，「新しいパソコンについている電気を節約する機能」などと表現することは容易に想像できる．利用者層を拡大するためには，このような問題を解決する必要がある．

3) 未知語登録

コンピュータ業界においては，新しい技術の開発とともに新しい用語(未知語)が生み出される．また，すでにある用語でも，技術の発展によりその意味が変化することも少なくない．このような技術の移り変わりが激しい業界においては，これらへの迅速な対応が不可欠である．この解決策としては，Mailing Listなどで配信されるコンピュータ関連の電子ニュースなどを使用した未知語の抽出が考えられる．

謝辞

最後に新Terminology作成にあたりご助力いただいた東京大学情報基盤センター図書館電子化研究部門の皆様およびマイクロソフト アジア リミテッドの矢部和博氏，森山永一氏，池辺聡氏，沼口繁氏，金起成氏，佐藤美智代氏，国分美宏氏，鹿野雄嗣氏に心より感謝する．

参考文献

[1] EDR電子化辞書 : http://www.iijnet.or.jp/edr/

[2] NTTコミュニケーション科学研究所監修，日本語語彙体系全5巻，岩波書店，1997.

[3] 情報処理振興事業協会，計算機用日本語基本名詞辞書IPAL : http://www.ipa.go.jp/STC/NIHONGO/IPAL/ipal.html

[4] 株式会社アスキー，アスキーデジタル用語辞典 : http://www.ascii.co.jp/ghelp/

[5] 株式会社インセプト，通信･情報辞典 e-Words : http://www.e-words.ne.jp

[6] NTT 出版，パソコン用語知ったか辞典 : http://www.nttpub.co.jp/paso/

[7] 羽田ゆかり，松本裕治，複数辞書の統合的利用のための汎用日本語辞書の構築，*情報処理学会研究報告*, **96-NL-116**, 1-6 (1996).

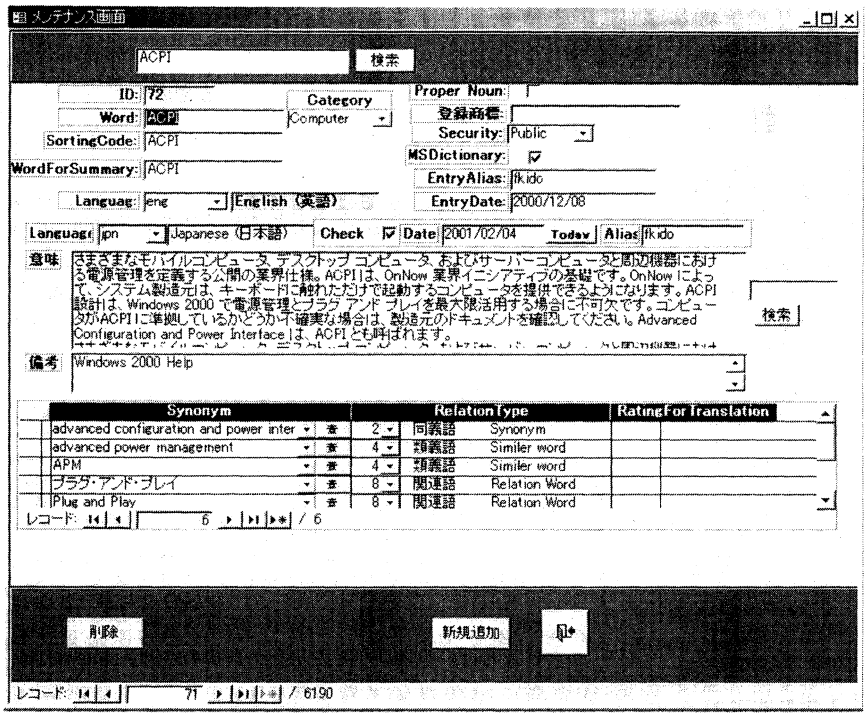

図 1. 辞書のメンテナンス画面

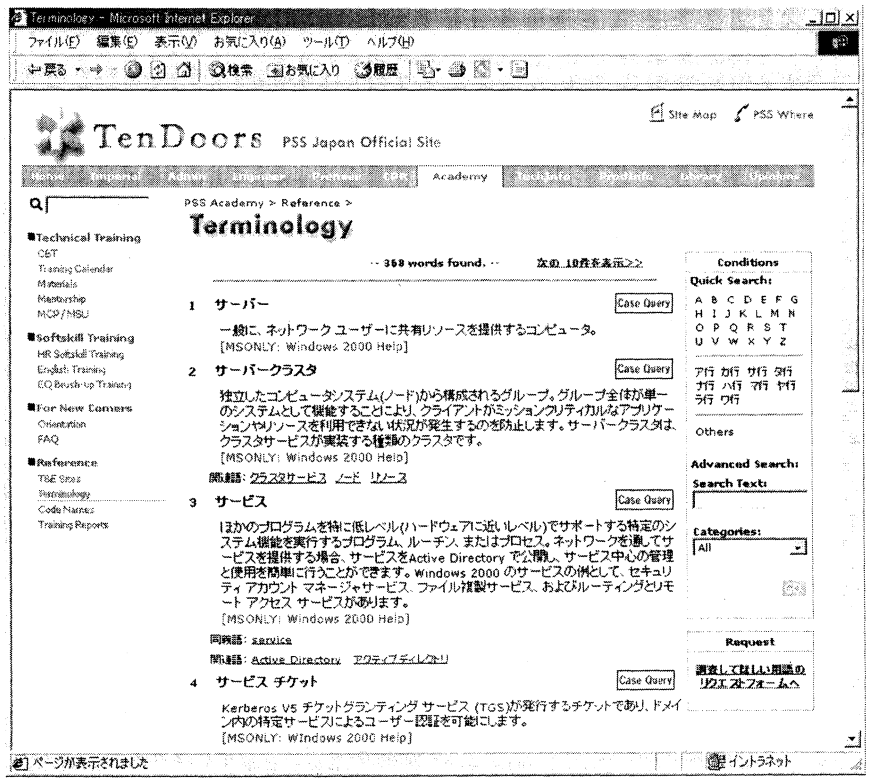

図 2. 社内公開した新 Terminology の画面