整合性の高い機械翻訳システム用辞書の構築法

葉茂 貴志　　フランシス ボンド
日本電信電話株式会社 NTTコミュニケーション科学基礎研究所
{hamo, bond}@cslab.kecl.ntt.co.jp

1 はじめに

自然言語処理研究には，機械翻訳，多義解消，情報検索など，様々な分野が存在する．一般的に各分野の研究には，専用の辞書が使用されるが，言語資源を有効活用するため，他の分野の研究にも用いられる汎用性の高い辞書が求められている．当研究所においても，機械翻訳システム用辞書(以下，辞書)が，多くの分野の研究に使用されている．しかし，現在の辞書は，データ形式がバイナリとなっているため，他の研究分野では扱い難い面がある．この問題点を解決するために，より汎用性の高い形式による，辞書のデータベース(以下，DB)化が求められている．

本稿では，現在の辞書を，RDBMS*1 を用いて再構築することで，幅広い分野で利用出来き，かつデータの一元管理およびデータ更新者情報の管理等における保守性の向上とヒューマンインタフェースの改善を図った辞書の提供を試みる．本稿では辞書のうち，主に英語辞書*2 を取り上げ，各品詞間でリンクを張った派生語*3 テーブルの作成によるデータの重複登録の防止，および参照整合性制約*4 の有効利用について述べる．

*1 Relational Data Base Management System
*2 機械翻訳システム用辞書は，①日本語辞書②日英対照辞書③英語辞書④パターン対辞書から構成されている．また，英語辞書は，①一般名詞②動詞③形容詞など，8つの品詞辞書から構成されている．
*3 見出し語の派生形 (例：名詞 base に対する形容詞派生語 basal)

2 現状

当研究所の機械翻訳システム用辞書[1][2]は，現在バイナリ形式で管理している．当初バイナリ形式を採用した理由は，読み出しの速度が速く，データ容量が少量で済むためであった．しかし，近年の情報技術の発展により，当時に比べると計算機の処理能力が数段向上したため，次の課題はメンテナンス性，および操作性に移行している．

バイナリ形式は，そのままの状態でデータを扱うのは困難であるため，1度テキスト形式に変更する必要がある．当研究所では専用の変換ツールを作成し，それを用いてテキスト化し，研究材料としてデータを扱っている．また，追加・削除等データを更新する際は，専用の変換ツールを用いてテキスト形式をバイナリ形式に戻す必要がある．これらの事が，他の研究分野では扱い難い要因の1つになっていると言える。

また，管理用ツールを採用していないため，同一データの二重登録防止の対応がされていないことや，テーブル化 (コード化) が十分にされていないため，1つの見出し語に対し1つの派生語しか登録する事が出来ないと言う問題点がある．

*4 テーブルの中の，1つまたは複数のフィールドに対して定義される規則．そのフィールドに入力される値が，関連した他のテーブルのフィールドから参照される値に依存しているときに，その値と一致する値のみを，挿入／更新する事が出来る．

3　検討

3.1　データベース化

これまで述べたように，現在の状態ではデータの管理体制に問題があり，時間を掛けて更新したデータが無駄になる可能性がある．また，バイナリ形式のデータは，操作性に問題があるため，扱いやすいデータに変更する事が望ましい．そこで，ファイルでの管理形式をＤＢ化し，汎用性が高く，かつ支援プログラムが充実している RDB（RDBMS）を採用する事にした．

DB 化[3]に伴い，以下のメリットが考えられる．

- 参照整合性制約を使用し，データの関連付けを容易に行える事が出来る．
- RDBMS を用いて DB を管理する事により，データ更新時にデグレード等のトラブルが発生しない様，排他制御を掛ける事が出来る．

3.2　派生語テーブルの導入

英語辞書は８つの品詞辞書から構成されており，そのうち一般名詞，動詞，形容詞，副詞の各品詞辞書には，その品詞が他の品詞に派生する場合，例えば次のように派生語を登録するエリアを設けている．

ｓｔａｒ	==>	ｓｔａｒｌｉｋｅ
(名詞)		(形容詞派生)
ｓｔａｒｒｙ	==>	ｓｔａｒ
(形容詞)		(名詞派生)

現在は，仮に派生元の「見出し語」が変更された場合，それに合わせて派生先の「見出し語」も変更する必要がある．しかし，リンクが張られていないため，片方のデータのみが更新されると言うような，人為的ミスが起こる可能性がある．

そこで「派生語テーブル」を設け，各品詞間の派生語に対しリンクを張る事により，人為的ミスを防止する．さらに，参照整合性制約の設定により，不整合な事象の発生を防止する．

4　実験

4.1　データベース化

・データベースの選択

辞書を，バイナリ形式から DB 化するにあたり，数多くある DB ソフトから，最適なものを選択する必要がある．その中から，PostgreSQL[4]を選択した．その理由を以下に述べる．

- PC UNIX 対応
  PostgreSQL は，ほとんどの UNIX およびUNIX互換プラットフォーム上で，動作している事が確認されている．
- 無償提供
  PostgreSQL は，企業，個人を問わず，無償にて利用する事が可能である．
- Unicode(UTF-8)対応
  多言語間での処理が可能である[5]．
- 参照整合性制約 対応
  「派生語テーブル」の作成，および資源の有効利用を達成するために，必要不可欠な条件である．

・データベース設計

現在のレコード仕様は，正規化，およびコード化が十分にされておらず，主キーとなる「見出し語」にも ID が付与されていない．派生語がある場合は，派生先の「見出し語」が直接登録されている．この件については 3.2 節でも記述した通り，データの不整合が発生する恐れがあるため，「見出し語」に ID を付与し，派生語テーブルの見出し語 ID を登録する事にする．

表示部*5についてはコード化され，IDが登録されているが，その部分はテーブル化されておらず，IDをそのまま表示している．IDのみを表示しても，システム利用者は意味が解らないため，この部分をテーブル化し，「内容」を表示するように改善する．

また，レコード長の保持など，プログラムに依存したデータ構造が取られているが，RDBの場合は，システム側で自動計算するため，この項目はDB設計の際には，意識する必要が無くなる．

・参照整合性制約

派生元のレコードが削除された場合，派生先とのリンクが切れてしまい，派生語テーブルに不要なレコードが残るという問題が発生する．このような事象が発生しないように参照整合性制約[5]を設定し，不整合の防止を図る．

・バイナリ形式への戻し

現行の辞書は，既に自然言語処理システムで使用されている．そこで，DBをバイナリ形式へ戻すツールを作成し，新しい辞書を従来のシステムでも直ぐ使えるようにする．

4.2 派生語テーブルの設計

派生語は，派生元となる各品詞辞書に登録されている．したがって，各品詞辞書から見出し語と派生語を抽出し，派生語テーブルを作成する．

5 効果

5.1 データベース

ログから更新者情報を取得する事が可能となり，仮に更新データに不備がある場合は，その原因を更新者情報から追跡する事が可能となった．

*5 見出し語に関する情報．母音種別，人称／性種別，複数形所有格形種別，語形変化語義依存，見出し語形で構成される．

5.2 派生語テーブル

派生語テーブルを設けた事による効果を以下に記述する．

(1)ある見出し語に対して1つの派生語しか登録出来なかった（もしくは，別レコードとして登録をしていた）が，1つの見出し語に対して複数の派生語を登録する事が可能となり，複合語生成の分野において有効な情報となった[6]．

(2)派生元の見出し語に変更が発生した場合，登録されたレコード数分の変更が必要であったが，1回の変更で済むようになり，操作手順の軽減化が図られた．

(3)“見出し語と派生語”の関係は，派生先の品詞辞書では“派生語と見出し語”として登録されているはずである．そこで，正しく登録されているか確認したところ，図1のように，片方の辞書にのみ登録している“見出し語と派生語”が存在する事が明らかとなった．

これまで片方向にのみリンクが張られていた派生語情報を，派生語テーブルを作成する事により，辞書を拡充する事が出来た．また，重複していた部分については，1つにまとめる事により，データ資源の軽減化を図る事が出来た．

6 今後の課題

- 現在の見出し語はUS表記だが，UK表記も表示出来るように対応する．
- 表示部について，現在日本語表記のみを表示する事にしているが，英語表記も表示するように対応する．
- RDBMSに対応した，辞書検索用ツールを作成する．

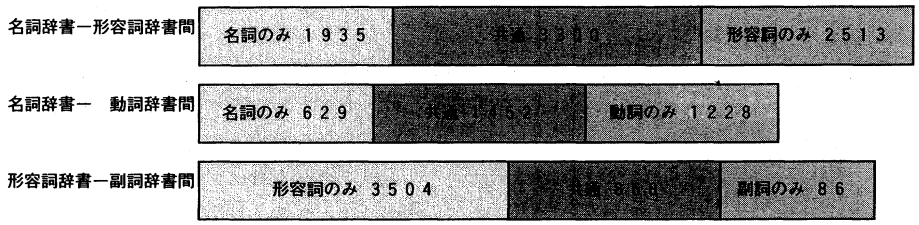

図1　品詞辞書間における派生語の登録件数

- 操作性,整合性を考慮したメンテナンス用ツールを作成し,随時更新が行えるようヒューマンインタフェースの向上を図る.
- 日英対照辞書や Word Net 等他の英語辞書ともリンクを張り,各辞書個別のデータを共有する事による情報量の拡大を試みる.
- XML 対応も視野に入れ,当システムに採用可能かを調査する.

7　おわりに

機械翻訳用システムを実装するには複数の辞書を所有する必要があり,設計当初は各辞書間の情報を共有していたが,各辞書個別の研究が進むにつれ,独自の情報が増加するようになる.辞書を整理し,各辞書の情報を相互間でリンクする事により,研究素材として頑強な DB と成り得る事がわかった.辞書を構築する事も大切であるが,メンテナンスを考慮した辞書作りと,稼動中の辞書を効率良く改造し,多目的の研究にも使用されやすくする事を心掛ける必要があるだろ.

参考文献

[1] 池原悟,宮崎正弘,白井論,横尾昭男,中岩浩巳,小倉健太郎,大山芳史,林良彦:日本語語彙大系,岩波書店 (1997)

[2] Satoru Ikehara, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa : Toward an MT System without Pre-Editing -- Effect of New Methods in ALT-J/E --, MT Summit III (1991)

[3] 鈴木昭男:実践!!データベース設計バイブル,ソフト・リサーチ・センター(1999)

[4] 石井達夫:PostgreSQL 完全攻略ガイド改訂第3版,技術評論社 (2001)

[5] 堀一成,前田彩,石島悌:多言語データベース検索アプリケーションの構築,情報科学技術フォーラム (2002)

[6] Takaaki Tanaka : Measuring the Similarity between Compound Nouns in Different Languages Using Non-Parallel Corpora, COLING-2002 (2002)