7H−7　データ標準化ツール(DBprompt/NAME)における複合語解析を用いた用語辞書構築方法

黒川 清　中川 優　関根 純
ＮＴＴ情報通信網研究所

1．はじめに

近年、社内で蓄積されている膨大なデータベースを共通資源として利活用することへの要求が高まっている。しかし、現状ではデータ項目の所在がわからない、データ項目の名称がわからない、名称が同じでも内容が同じである保証はない、といった問題がある。そこで、我々は、データ項目を分かりやすい名称に統一し、さらにデータ項目の属性、桁数の統一を図るデータ標準化の重要性を認識し、それを支援するツール(DBprompt/NAME)[4]を開発した。

NAMEは、既存システムのデータ項目の所在を明らかにするために、似かよった名称のデータ項目を検索する類似検索機能、データ項目生成時に分かりやすく統一的な名称をつけるために、Durellの命名規則[1]（図１）に違反した名称をチェックする機能を持つ。これらの機能の核となるのは、データ項目名称を構成する用語を管理している、用語辞書である。しかし、これまでは、

①用語辞書構築時に必要な用語区切り

②名称チェック時に使用される用語種別（図１）

の設定基準が明確でなかったため、用語辞書に不備が生じ、それを用いた名称チェックなどの処理の信頼性に問題があった。

本報告では、データ項目名称が複合語であることに着目し、その構成要素である品詞を用いた用語区切り、品詞と意味カテゴリを用いた用語種別の判定法を、それぞれ２章、３章で述べる。また、４章ではこれらの判定法を用い、既存の用語辞書原本を評価した結果について述べる。

2．用語区切り判定法

Durellの命名規則においては、データ項目名称に用いる用語を、その一語で、事物、事象、値など、意味を持つものとしている。一方、品詞は名詞、動詞、形容詞など、様々なものを含むとともに、接辞である「者」、「所」のように、それだけでは意味を持たない付属語がある。このような事実のもと、

・データ項目に使用される品詞は限定される

・用語は一つ以上の品詞の組み合わせにより構成されるという考えにより、特定分野の１１システム約１万データ項目に出現する、用語と品詞の関係を整理した結果、表１に示すような用語構成を得た。

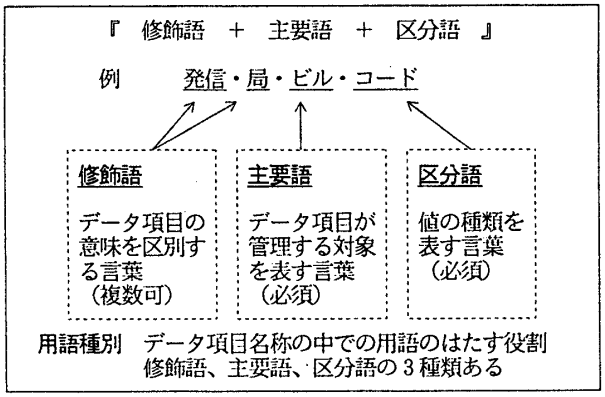

図１　Durellの命名規則

品詞の組み合わせによる用語区切り判定法の有効性を確かめるため検証実験を行った。実験では、上記とは別のシステムのデータ項目の名称を、NTT研究所開発のキーワード自動抽出システム(INDEXER[2])で複合語解析処理し品詞毎に分解して、表１に従い用語を生成した。その結果、330項目から生成された888用語のうち、97%の用語が正しく生成された。区切り誤り(3%)の原因は、データ項目名称にNTT特有の略号が使用されているものがあったり、INDEXERの日本語辞書に複合語が登録されているため、正しく処理できなかった。これらは、INDEXERの日本語辞書を改善することにより解決することができるので、この手法の有効性は確認できた。

3．用語種別判定法

文法上の役割によって分類したものが品詞であることから、名称内の用語の果たす役割をあらわす用語種別の判定に品詞を用いることを検討した。２章と同じ、約１万データ項目に出現する、品詞と用語種別の関係を調べた結果、品詞だけで用語種別が判定できるものと、品詞だけでは判定できないものがあることがわかった。

3．1　品詞による用語種別判定

ＮＴＴ社内で用いられる用語のうち、一般名詞と時詞を除く他のものは、既存の用語辞書原本(1296用語)において約４割を占めるが、これらは品詞の種類により用語種別が決まることがわかった。ただし、接尾辞は種類が少ないので、ヒューリスティクに用語種別を決定した。用語種別の判定は、以下のように考えた。

表１　品詞の組み合わせとその例

品詞	用語の例
一般名詞	上位、回線、番号
形動語幹	固有、可能、異常
時詞	月間、休日、午後
副詞	以上、随時
サ変名詞	受注、契約、修正
固有名詞	ＮＴＴ
他動詞	引込、打切、振込
動詞転生型	受付、繰越、割当
接尾辞	値、数、量
接頭辞＋一般名詞	最下位、上支線
接頭辞＋サ変名詞	無応答、大代表
接頭辞＋動詞転生型	再割当
接頭辞＋形動語幹	不完全
一般名詞＋接尾辞	自動的、事業部
サ変名詞＋接尾辞	契約者、交換所
他動詞＋接尾辞	振込済
動詞転生型＋接尾辞	受付後
数詞＋助数詞	第一種、１．５Ｍ

On building thesaurus for data standardization tool (DBprompt/NAME) using compound word analysis
Kiyoshi KUROKAWA, Masaru NAKAGAWA, Jun SEKINE
Network Information Systems Labs., NTT

○動詞転生型、サ変動詞型、他動詞名詞
・動作を表すもので、イベントの事象を表すと考えられるので主要語とした。
○形容動詞語幹、固有名詞、数詞＋助数詞、副詞
・名詞に接続してその名詞の意味を限定するものなので修飾語とした。
○接辞
・基本的には単独では用語となりえないが、「名」、「値」等のように、データ項目に設定される値を表すものは区分語とした。
・２つの用語にかかる場合は、単独で修飾語とした。
例：公衆＿専用＿別
・接辞が他品詞と接続した場合、その品詞の用語種別を継承する。　ただし、接尾辞の「的」、「用」が接続した品詞については、用語種別が修飾語に変化する。

品詞と用語種別の関係を整理した結果を表２に示す。

品詞を用いた用語種別判定法の有効性を確かめるために、あるシステムのデータ項目名称により検証実験を行った。生成された888用語のうち、品詞により用語種別が判定できるものは443用語あり、これらは100%正しく判定でき、品詞を用いた用語種別判定法の有効性が確かめられた。

３．２　意味カテゴリによる用語種別判定

既存の用語辞書に登録されている用語の約６割を占める、一般名詞、時詞については、品詞だけでは用語種別の判定はできないので、INDEXERが出力する意味カテゴリを用いて、さらに判定を行うことにした。意味カテゴリとは、語が持つ意味の基本概念である[3]。意味カテゴリに基づく用語種別判定は、以下のように考えた。

○具体的な事物を示すものは、名称における主語になると考えることができるので主要語とした。
○具体的な事物を表すもののうち、値を表す特殊なもの（番号、名前など）を区分語とした。
○具体的な事物を修飾するものは、対象を伴い関係などを表すので修飾語とした。

なお、INDEXERの処理の結果、３つの意味カテゴリが出力されるので、第一候補を用いて用語種別を決定した（図２）。

意味カテゴリを用いた用語種別判定法の有効性を確かめるために、検証実験を行った。生成された888用語のうち、意味カテゴリにより用語種別が判定できるものは445用語あり、89%正しく判定できた。誤りは11%であったが、その原因は、INDEXERが第一候補として出力したものが、データ項目で使用される意味と異なっていたためである。しかし、出力される３つの意味カテゴリに対して、適切なものを選択できれば用語種別判定の精度は向上すると思われる。よって、意味カテゴリを用いた用語種別の判定法の有効性は確かめられた。

４．用語辞書の検証

２章、３章の用語区切り、用語種別の判定法を用いて、現在構築されている用語辞書原本(1296用語)の見直しを行った。その結果、用語辞書原本の用語区切り、用語種別の設定誤りが計17%あった。先行して、既存の用語辞書原本を用いて約１万データ項目の名称チェックを行ったが、用語辞書の誤りのためにエラーとなった名称は、エラーの4.5%を占めた。用語辞書の誤りが名称チェックに及ぼす影響は潜在的なものであり、実際の用語辞書の誤りは約４倍であることがわかった。これらを修正することにより、用語辞書原本の品質も向上した。

５．まとめ

本報告では、データ標準化のための用語辞書品質向上手法を述べた。この提案により、非専門家でも用語辞書の構築が容易に行えるようになり、データ標準化作業が進展するようになった。

今後の課題は、用語種別判定の精度向上のために、複数の意味カテゴリ候補から最適な候補を選択する方法を検討することである。

参考文献）
1)W.R.Durell：ﾃﾞｰﾀ資源管理,日経ﾏｸﾞﾛｳﾋﾙ社(1987)
2)木本晴夫：日本語新聞記事からのキーワード自動抽出と重要度評価,信学論,Vol.J74-D-Ⅰ,No.8(1991)
3)国立国語研究所：分類語彙表,秀英出版(1964)
4)関根他：ﾈｰﾐﾝｸﾞ手法と支援ﾂｰﾙ,信学技報,DE89-4(1989)

表2　品詞と用語種別の関係

品詞 / 種別	一般名詞	時詞	固有名詞	サ変動詞型名詞	動詞転生型名詞	他動詞	形容動詞型名詞	数詞＋助数詞	副詞	接尾辞	他品詞＋「用、的」
修	上位	月間	NTT				異常	1.5M	以上	間、別	営業的
修主	電話	休日		登録	受付	振込					
修主区	番号	曜日								名、値	

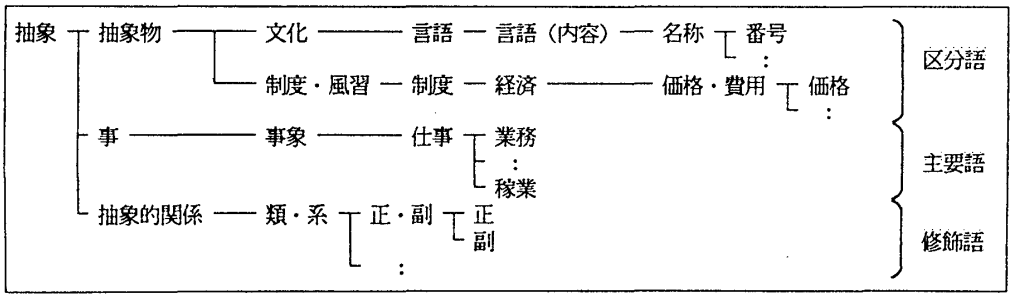

図2　意味カテゴリと用語種別（抜粋）