5B-4

専門用語辞書の試作概要

熊野 明 須之内美幸 石井 敬子
(株)日本電子化辞書研究所

1. はじめに

(株)日本電子化辞書研究所(EDR)では，自然言語処理用の大規模な電子化辞書の研究開発を行っている[1]．昭和62年度には，単語辞書の一部として専門用語辞書に関する研究開発を開始した．

本論文では，専門用語辞書の試作における方針，開発内容，問題点とその対策を報告する．

2. 研究開発方針

EDR辞書は，日本語と英語の見出し語のもつ全ての意味を概念単位で整理したものである[2]．

専門用語辞書は，EDR辞書の中で単語辞書の一部を構成し，基本語辞書と対をなしている(図1)．基本語辞書には日本語，英語各20万語について情報を記述してある．しかしその語は分野によらない共通のものであり，実際に専門分野の文章を処理しようとする際には語彙が不十分である．専門用語辞書はこれを補う役割を担っており，現実の文章処理に必要な語を積極的に収録している．

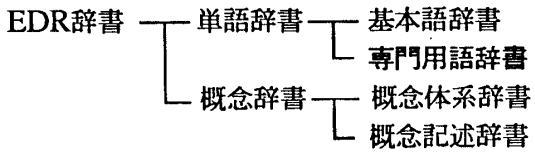

図1. EDR辞書の構成

一般に専門用語辞書としては，これまでに多くの機関によって整備されたものが利用されている．しかしこれらのあるものは語彙の範囲を限定して何らかの標準規格として設定したものであり，またあるものは見出し語と対訳のみを大規模に揃えたものである．いずれも意味処理まで考慮した自然言語処理を実現するためには，量的にも質的にも十分とはいえない．EDRの専門用語辞書は計算機による自然言語処理を考慮した量と質を備えるものである．

対象とする専門分野としては，「情報処理」を扱っている．語彙数は，日本語，英語ともに5万語の規模で第一次試作中である．

3. 見出し語

3.1 見出し語選定

専門用語辞書を作成する際，見出し語の選定は最大の課題のひとつである．見出し語と認めて辞書に収録すべきか否かの判断は，困難を伴う場合が多い．殊に情報処理分野では問題が大きい．

なぜなら，専門用語としての語彙の範囲が急速に広がっているからである．対象とする情報処理分野では数多くの論文，書籍，雑誌が頻繁に出版され，その文章中には新たに作られた専門用語が含まれていることがある．

例えば，既存の概念を表す用語を複数組合せて，新たに必要とされる概念を表す用語が作られる．専門用語ではこのように複合語によって新しい用語を形成する例が極めて多い．また，海外の文献に現れた外国語を翻訳した結果として，あるいはそのまま外来語表現として新たな用語が作られる．情報処理分野は国際的な分野であるため，外来語が翻訳されないまま一般に使用され，そのまま用語として確立する例も多い．さらに，従来は基本語として用いられている語を借用して，専門用語の一つとして認めることもある．専門分野で用いる新しい意味と基本語がもつ本来の意味が類似している場合で，基本語に専門分野の意味を割当てる手法である．このように情報処理分野の専門用語は，日々その数を増しているといっても過言ではない．

Development of a Technical Term Dictionary for Natural Language Processing
Akira KUMANO, Miyuki SUNOUCHI and Noriko ISHII
Japan Electronic Dictionary Research Institute, Ltd.

そこでEDRでは，専門用語辞書の見出し語選定基準を設定した．具体的には，情報処理分野の文献に現れる頻度の高い語彙を選択する．基本語辞書に含まれているものでも専門用語として特定の意味をもつものは選定する．その上で特定の専門分野に偏らないよう，また，対になる語がもれないよう全体のバランスも考慮する．

3.2 正表記

見出し語を整理する際，もう一つの問題は，その表記法である．専門用語を実際の文章中で検索すると，表記の多様性が認められる．日本語の場合，漢字の送りがなや外来語のカタカナ表記(中黒，長音，促音)に多様性が存在する(例1)．英語の場合，本来複合語である用語を複数単語のまま表記するか，ハイフンで連結するか一単語化するか等，一般には定式化されていない点が多い(例2)．これらの性質は基本語の場合と特徴の異なる場合もある．

(例1) プロセサ，プロセッサ，プロセッサー

(例2) data base, data-base, database

EDRでは上の状況を整理し，見出し語の正表記基準を設定した．この基準に適合するものを正表記として登録し，その他の表現は全て異表記として辞書中に記述する．ところが日本語で特にカタカナ語が複合語を構成すると，その異表記の形態は組合せ的に増加する．そこで辞書中には多様な形態を規則的に整理した形式で記述する．必要に応じて可能性のある異表記を個々に生成することによって実際の文章に現れる多様な表現に対応できる構成にしている．

4. 記述形式

辞書中のデータ構造は基本語辞書と同様である．見出し語情報，品詞情報，概念別情報，訳語情報が，階層的に構成されている(図2)．

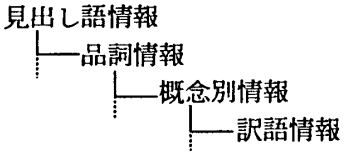

図2. EDR辞書のデータ構造

基本語の場合，一つの見出し語に複数の概念が存在する場合が多いが，専門用語の場合，ほとんどの場合が，見出し語一つに対して単一の概念である．

5. 対 訳

専門用語としての性質上，日本語の概念と英語の概念はほとんど一対一に対応する．基本語辞書で多くの場合に一対一に対応せず，日本語の概念と英語の概念の間を「上位」,「下位」等の関係で示す必要が生じたのと状況を異にしている[3]．これは，専門用語の概念は日本語，英語独立に形成されたものではなく共通のものだからである．それを表す用語は言語表現よりも「記号」に近いものである．ある専門用語概念が形成された時点でその言語で記号として表現され，他の言語に導入されるときには，そのまま概念と用語が対応している．

ところが，英語には用語が存在するが日本語には日本語としての用語が確立していない場合がある．このほとんどの場合，英語世界で確立した概念に与えられた英語表現の用語をそのまま日本語で使用している．日本語の対訳を与える必要性が認められるまでに達していない場合である．この場合，適切と判断できる日本語の用語を割当てることができれば，訳語情報として記述する．

6. おわりに

これまで述べたように，EDRでは情報処理の分野に関する専門用語辞書を試作中である．今までにない規模での自然言語処理用専門用語辞書であるため，目標の量と質を実現するまでには，現段階では顕在化していない問題も起こりうるであろう．EDR辞書が広く標準として利用されるよう研究開発を継続していく．

参考文献

[1] 河田他:「大規模自然言語処理用電子化辞書の開発」, 情報処理学会第34回全国大会，1X-7 (1987).

[2] 内田他:「自然言語処理のための電子化辞書の構成法」, 情報処理学会第35回全国大会，1S-5 (1987).

[3] 熊野他:「自然言語処理のための対訳辞書における訳語選定」, 情報処理学会第35回全国大会，1S-3 (1987).