7S−3

ニュース専門用語英日辞書の構築

畑田のぶ子*　相沢輝昭*　木村展幸**
(　*NHK放送技術研究所　**漢字情報サービス　)

1. はじめに

NHKでは、現在衛星放送ワールドニュースで、試験的に英日機械翻訳システムを使用し、テロップの作成をおこなっている。放送局のような報道機関ではニュースの対象範囲は広く、必然的に辞書の語数は多くしなければならない。

今回は、冊子型辞書データを用いて構築したニュース用語辞書（暫定版）について報告する。

2. ニュース用語辞書（専門用語辞書）の役割

われわれの英日機械翻訳システムは次の2種類の辞書を使用している。

基本辞書:

基本的な語を登録した辞書で、見出し語、品詞、ウエイト、品詞細分類、内在素性、訳語からなる。

オプション辞書:

専門用語や分野別に用語を登録する辞書で、見出し語、品詞、訳語からなる。簡単な登録で基本辞書より優先的に用語を翻訳に反映させることができ、このような辞書を複数個用意し、優先順位をつけて組み合わせて利用することが可能である。

ニュース用語辞書（専門用語辞書）は、このオプション辞書にニュース用語を登録したものである。

ニュース用語の収集にあたっては各種専門辞書、外電などの利用が考えられるが、今回の構築にあたっては下記のものをソースデータとして使用し、それぞれを個別の専門用語辞書とした。

辞書名称	ソースデータ
専門辞書1	三省堂ニューズ英語辞典
専門辞書2	小学館最新英語情報辞典（第2版）
専門辞書3	集英社・IMIDAS（1988）

以下ニュース用語辞書を専門辞書と呼ぶ。

3. 構築方針

今回利用したソースデータは基本的に冊子型である。それを人手または計算機により定型のフォーマットに変換し、辞書作成用の入力データとした。このようなデータをすべてチェックしてから辞書作成を行うと非常に時間がかかり、また問題点を洗い出すのが難しい。そのため、次のようなステップで辞書の作成をおこなった。

ステップ1:

計算機で機械的に処理可能なデータのみを利用して暫定版の辞書を作成する。処理不可能だったデータの問題点、暫定版作成時に明らかになったデータ上の問題点を検討する。

暫定版の専門辞書を用いて翻訳テストを行い、専門辞書としての問題点を明らかにする。

ステップ2:

ステップ1で明らかになった問題点を基に辞書作成用の入力データ、辞書作成プログラムを修正する。

その入力データとプログラムを用い、本番用専門辞書を作成する。

ステップ1、ステップ2の関係を図1に示す。

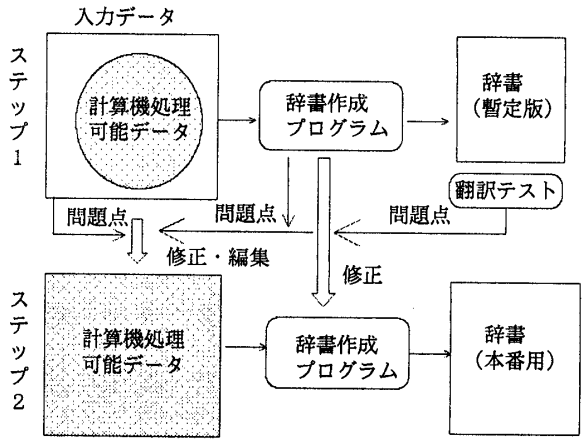

図1

4. ニュース用語辞書の作成（暫定版）

辞書作成用データの内容

項目:（見出し語、品詞、訳語、備考）

見出し語: 普通見出し語、複合語、成句、慣用句、略語

品詞: 複合語、成句、慣用句には品詞の指定がない

訳語: 複数の訳語を含む

備考: 名詞・普通見出し語の複数型（特別な場合）、動詞の不規則変化、分野、ユーセイジ（記述形式は不定型）

計算機で一意に編集可能なものは処理可能データとした。また、入力データからできる限り複合語も採用するようにした。

計算機処理可能データとして設定した条件、およびデータ生成条件は下記の通りである。

（1）見出し語の検討

見出し語の中で次のものを含む語は除く

・言い換え、異綴り、省略記号を含む見出し語
（例:　[・・・], ", " (・・・)）
・1文字の見出し語
・前置詞ではじまる見出し語

Construction of special MT dictionary for news English

by N.Hatada(1), T.Aizawa(1), N.Kimura(2)
(1)NHK, (2)Kanji Information Service

・人称代名詞を含む見出し語
・one's person'sなどを含む見出し語

（２）品詞の検討
・名詞：見出し語につけられた品詞が、名詞のみのもの、品詞の指定無しのものを名詞用語の候補とする
但し、指定無しのものの中で名詞でないと判断したものは除く
・副詞：見出し語につけられた品詞が、副詞のみのものを副詞用語の候補とする
・動詞：見出し語につけられた品詞が、動詞のみのものを動詞用語の候補とする
但し、訳語の語尾チェックにより訳語エラーと判断したものは除く
・形容詞：見出し語につけられた品詞が、形容詞のみのものを形容詞用語の候補とする
但し、訳語の語尾チェックにより訳語エラーと判断したものは除く

（３）訳語の検討
・訳語データのないものは除く
・参照記号のあるもの、省略記号のあるものは除く（例：＝，－＞、．．．）
・言い換えデータは除く（例：［...］　）
・英大文字、／、＆のみで構成される見出し語は略語とし、訳語は見出し語と同じとする
（訳語がフルスペルで記述されていることと、数種の訳が考えられるため）
・形容詞の訳語で修正の必要のあるものは可能な限り計算機で自動修正する。自動修正不可能なものは要チェックとして洗い出す
（例：頭のいい--＞頭がよい、
嵐のふきまくる：修正の必要があるが、語尾による自動修正不可能）

（４）備考：備考のデータは暫定版では利用しない（形式が不統一）

（５）基本辞書との検討
基本辞書にある語は除く

表１：辞書ソースデータと専門用語辞書（暫定版）作成のデータ内訳

	ソース見出し語数	処理不可能データ（%）	処理可能データ数	基本M辞書とのダブリ語数
専門辞書１	48,253	18.49%	39,330	11,520
専門辞書２	31,750	26.81%	23,238	2,965
専門辞書３	9,686	2.20%	9,473	1,102
合計	89,689	19.68%	72,041	15,587

表２：専門用語辞書（暫定版）見出し語数・見出し語の形式

	品詞						大文字始	全大文字	ハイフン含む
	名詞	副詞	形容詞	自動詞	他動詞	合計			
専門辞書１	21,983	481	2,903	380	1,010	26,757	3,244	841	1,368
専門辞書２	18,422	69	1,186	189	296	20,162	4,719	2,107	1,690
専門辞書３	8,116	0	0	0	0	8,116	2,246	8	679
合計	48,521	550	4,089	569	1,306	55,035	10,209	2,956	3,737

４．辞書（暫定版）作成結果

辞書作成のための入力データの内訳を表１に、作成した辞書の内訳を表２に示す。また、専門辞書間のダブリの状態を図２に示す。

専門辞書の８８％は名詞である。

専門辞書１では基本辞書とのダブリも多く、名詞以外の品詞も多い。これは三省堂ニューズ英語辞典が一般的な用語も多く含んでいることによる。また、処理不可能とされたデータが多いのは、慣用句、成句を多く含むためである。

専門辞書２には略語、大文字で始まる語が多い。これは小学館最新英語情報辞典が略語、書籍・雑誌名、ミュージカル・歌の題名、機関・組織名を多く含むことによる。処理不可能とされたデータが多いのは、特別な氏名、説明的な訳語を含むためである。

集英社・ＩＭＩＤＡＳは名詞のみなので専門辞書３はすべて名詞である。

５．今後の課題

今回の構築で明らかに成った課題は次の通りである。

（１）処理対象外になつたデータについても利用する必要がある

（２）略語、人名、組織名、国名が多く含まれるが、これらは別々に専門用語辞書とし、局所処理などに利用するほうが辞書管理の面でもよい。

（３）訳語の優先順位を検討し、分野設定が必要かどうかを検討する必要がある。

（４）基本辞書とダブリの語、動詞、形容詞については、基本辞書への組み込みを検討する必要がある。

今後、翻訳テストと合わせて、以上の点を検討し、本使用のニュース用語辞書を構築する予定である。

専門用語辞書・ダブリの状態（語数）

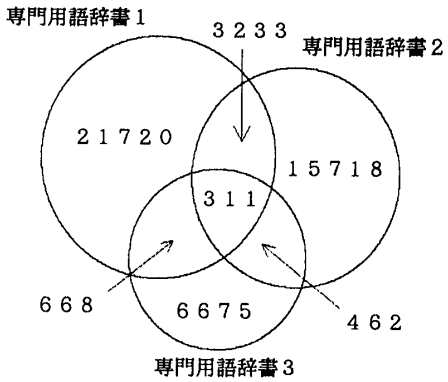

図２