7S−2

英日機械翻訳のための基本語辞書の拡張

田中英輝　*有賀憲和　*松田健生
NHK放送技術研究所　*カテナリソース研究所

1　はじめに

ＮＨＫでは，１９８９年８月に衛星放送英語ニュースのテロップ（字幕）作製に機械翻訳を導入し[1]，現在もその試用を続けている．この試用と並行して，翻訳率の向上を目指した，文法，インターフェース[2]，ニュース固有表現局所解析[3][4]，基本語辞書，オプション辞書[5]の改良をを続けている．本稿ではこの中の基本語辞書の改良について報告する．

我々のシステムの辞書は，基本語辞書とオプション辞書の２部で構成されている．オプション辞書は，名詞を中心としたニュース専門用語を記述するための辞書である．また，日々新しい単語が出現するニュース文翻訳に即応するため，簡単な辞書記述のみで翻訳に使用できるようになっている．

これに対して，基本語辞書はニュース文に限らない一般的な用語を記述するための辞書である．また，翻訳時に必要となる文型パターン，共起情報など，さまざまな情報を記述している．基本語辞書は，導入当時，約２万語の辞書エントリしかなく，翻訳率向上には，語彙の拡張が必要であった．一方，基本語辞書に記述されたさまざまな訳語を，適切に選択する戦略も必要となり，意味マーカーを利用した訳語選択の研究を開始した．以下具体的に我々のこれまでの研究について報告する．

2　基本語辞書の語彙追加

追加前の基本語辞書（以下，旧基本語辞書と略す．）の品詞内訳を表1.に示す．

表１．旧基本語辞書品詞内訳

総エントリ数　２０，７１５

品詞	個数
動詞類	４，４１９
助動詞類	５７
形容詞・冠詞類	５，４２５
副詞類	２，５２３
名詞類	１３，１８７
接続詞類	６６
疑問詞・関係詞類	２６
前置詞，その他	３８６

この辞書に対して，学習研究社から発売されている機械可読の英和辞典（以下学研辞書と略す．）の見出し語を追加した．

語彙の追加は，プログラムによる自動化を基本としたが，学研辞書は括弧，カンマ，コロンの使用法に不統一な部分があり，精密な情報抽出は断念した．そのかわり基本語辞書として必要な共起情報や文型パターン情報を，語義記述の中の［・・・を］や［～の］などに着目して，人手で追加した．

2．1　機械可読辞書の利用

旧基本語辞書にすでに採録されているエントリについては変更を行わず，採録されていない語彙のみを追加した．表２に，学研辞書の品詞内訳の一部と，我々の辞書になかった追加対象語彙の品詞内訳を示す．

表２．学研辞書内訳及び追加内訳

	学研辞書	追加対象数
名詞	２５，６４０	１５，９８７
形容詞	８，２７０	４，５９４
副詞	３，０４２	２，０４１
動詞	５，４５０	１，８８４

今回追加された語彙は，名詞，形容詞，副詞，動詞だけであり，その他の品詞は旧基本語辞書に登録済みであった．追加語彙は，名詞がその大半を占めている．

2．2　辞書エントリ増加の効果

拡張辞書と旧辞書の両方を用いた翻訳実験を行った．翻訳対象は，（１）１９８９年７月から１９９０年３月までに，実際にワールドニュース字幕作成のため，入力された文章５０２５文と（２）ＡＰ通信社の外電１９９０年１月分１０２４文である．表３にそれぞれ変化した文章の数を示す．

表３．辞書拡充による翻訳変化数

	ＡＰ電	ニュース
文章数	１，０２４	５，０２５
変化数	３５８	７７７

変化した数が，そのまま翻訳の改善につながるわけではないが，大まかに確認した感じでは，かなり改善された印象を受けた．

改善された例

1．New York already has mandatory classes to teach cabbies courtesy.
2．ニューヨークはcabbies儀礼を教えるために、義務的クラスをすでに持っている。
3．ニューヨークはタクシーの運転手儀礼を教えるために、義務的クラスをすでに持っている。

An Enhancement of *NHK*'s MT core dictionary
Hideki TANAKA(NHK) Norikazu ARUGA(CLI)
Takeo MATSUDA(CLI)

1．You might say he has gone goofy.
2．彼が過ぎさったgoofyを持っているとあなたは言うかもしれない。
3．彼がばかなようになったとあなたは言うかもしれない。

改善されなかった例

1．The system makes 3 dimensional images from the satellite pictures.
2．システムは、サテライト絵から3 dimensionalなイメージを作る。
3．ステムは、サテライト絵から3 *寸法*のイメージを作る。

1．Saint Louis
2．Saint Louis
3．*聖人名につける称号*Louis

1．Ken Kashihara remembers her career.
2．Ken Kashihara は、彼女の生涯を思いだす。
3．*理解*Kashihara は、彼女の生涯を思いだす。

1：原文、2：旧辞書での翻訳、3：新辞書での翻訳

改善されなかった主な原因は，（1）人名や，大文字で始まる語に特殊な意味があり，それが訳されたもの，（2）訳語というより，説明が出力されたもの，（3）訳語の選択が不適切なものの3つであった．

（1）に関して，われわれのシステムでは，人名は局所解析用の辞書で取扱うことになっており，基本語辞書からは削除する予定である．（2）に関しては，説明的な訳語を長さで検査し，12文字以上の訳語の確認は行ったが，短い訳語にまだ残っているようである．

我々のシステムでは，オプション辞書と基本語辞書の両方に登録された語がある場合，オプション辞書の訳語を優先して使用することになっている．そこで基本語辞書の中で，訳語が確定的な専門用語に近いものはオプション辞書に，オプション辞書の中で多くの訳があるものは基本辞書へ移動する，などの辞書間の調整が必要である．

3　訳語選択

辞書エントリの数の増加と並んで，翻訳率向上に欠かせないのが，訳語選択である．我々の翻訳システムでは，訳語の選択手段として意味マーカーを使うことができる．このため，階層的な意味マーカーを81個設定し，辞書への付与を進めている．

3．1　基本語辞書名詞への意味マーカー付与

名詞訳語への意味マーカー付与は，類語国語辞典[6]を用いて半自動的に行った[7]．現在，旧基本語辞書の名詞訳語への付与を終えたところである．表4に，この方法で付与された再下位階層の意味マーカーの頻度統計の上位を示す．

数多く付与されたマーカーは（1）CONACT，ABSACT，MONOACTなど，抽象名詞でかつ動詞の派生形（サ変名詞など）を持つのものに付与するもの，（2）NOWILLCHA，抽象名詞でかつ形容詞派生形を持つものに付与するもの，（3）HUMSTATIC，抽象名詞で人間にかかわる概念，などである．

具体名詞の中で多かったマーカーは，再下位マーカーではないが，HUMANが2000程度，ORGANIZATIONが800程度付与された．

表4．意味マーカー頻度統計（再下位階層）

意味マーカー	頻度
CONACT	2，397
ABSACT	2，324
HUMSTATIC	2，263
MONOACT	2，053
NOWILLCHA	1，357
INSTRUMENT	1，046

我々の意味マーカーは上記（1），（2）に示すように，抽象名詞の分類を派生形を手がかりに行った部分がある．このようなマーカーの使い道は，例えば，makeの目的語としてサ変動詞派生形を持つ訳語が共起した場合，動詞の訳を「する」にすることなどを考えている．

make　a　stop　=>　停止する
stop　（名）　停止［サ変動詞］

今回の意味マーカー付与ではこのようなマーカーが多用された．期待される訳し分けが行われるかどうかは今後調査したい．

4　おわりに

基本語辞書の拡張及び，意味マーカー付与について述べた．

新しく追加された語彙の辞書記述は，旧基本語辞書に比べて大まかである．詳細情報を自動処理でこれ以上追加することは無理だと思われるため，人手により行う予定である．

意味マーカーについては，大まかなものを旧基本語辞書名詞へ振り終えたところである．これから小規模な動詞辞書へ意味マーカー記述を行い，翻訳実験を行う予定である．

【参考文献】

（1）相沢，他：「衛星放送ワールドニュースの英日機械翻訳」情報処理学会第40回全国大会2F-1
（2）住吉，他：「機械翻訳系と放送系のインターフェイス」情報処理学会　第41回全国大会投稿予定
（3）加藤，他：「英日機械翻訳における固有名詞処理」情報処理学会　第40回全国大会2F-2
（4）加藤，他：「英語ニュース文におけるハイフンを含む語の局所解析」情報処理学会　第41回全国大会投稿予定
（5）畑田，他：「ニュース専門語英日辞書の構築」情報処理学会　第41回全国大会投稿予定
（6）大野，浜西：「類語国語辞典」角川書店
（7）田中，他：「類語国語辞典を介した意味マーカー付与」情報処理学会　第40回全国大会6F-7