日英機械翻訳用スケルトン－フレッシュ型構文意味辞書の構成

6Q－8　横尾昭男　中岩浩巳　白井諭　池原悟

ＮＴＴ情報通信網研究所

１．はじめに

筆者らは日英機械翻訳の研究を進めており，日本語の意味構造を解析して英語の意味構造に変換するために構文意味辞書を構築してきた。従来の辞書記述方式では，１つの意味表現構造として１つのパターンですべてを記述していたが，ここでは，そのパターン構造を骨格部分(スケルトン)と肉付部分(フレッシュ)に分離して辞書に格納する方式を提案する。

本方式により，既存の約13,000パターン*1を再構築した結果，構造の持つ意味の共通化を行うことができ，少数の骨格構造で大多数のパターンを記述できることを示す。

*1: 当初15,000パターン作成したが，その後縮退などにより，現状では約2,000パターン減少している。

２．構文意味辞書の構成

ＮＴＴで研究を進めている日英機械翻訳システムALT-J/Eでは，「構造は意味の一部である」という考え方に基づき，単位文を変換するときには，“多段変換方式”に従った構文意味辞書を構築してきた（図１）[1],[2]。

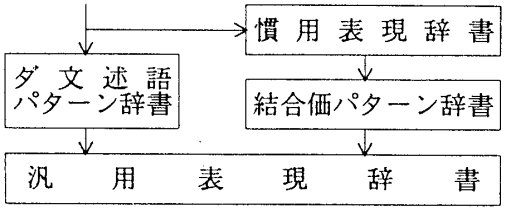

図１　多段変換用構文意味辞書

例として，「～が～を～に提案する」という文を変換する結合価パターンを図２に示す

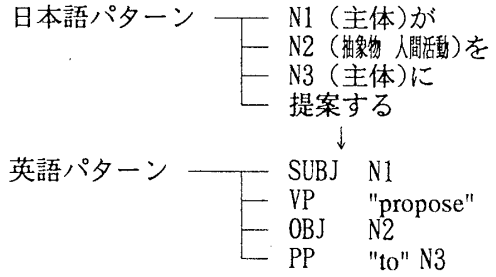

図２「提案する」の結合価パターン

この英語パターンは，全体を１つの構造とするので，全体を扱いやすいという利点はあるが，複雑な構造になると全体を捉えにくくなり，共通部分について重複記述する必要があるので，全体の記述量が増えるという問題がある。

３．スケルトン－フレッシュ型構文意味辞書

３．１　構文意味辞書の構成法

英語パターンの重複を考慮し，全体を効率よく記述するため，図２における英語パターンを**骨格部分**と**肉付部分**に分離して記述する方式を提案する。骨格部分と肉付部分は，それぞれ以下の特徴を持つ。

骨格部分(ｽｹﾙﾄﾝ) … 構造の意味を表す
- 述語，格要素（前置詞句も含む），副詞句，修飾句節の構造を表す
- 単語の品詞も記述する

肉付部分(ﾌﾚｯｼｭ) … 要素の意味を表す
- 格要素の変数(Nx)へのポインタ，単語の字面等を記述する

骨格部分は，図１における各辞書共通に１本作成し，それを**新英語パターン構造辞書**と呼ぶ。肉付部分は，各辞書ごとに対応づけて作成し，全体をまとめて**新英語パターン辞書**と呼ぶ。

図２における「提案する」に対応する骨格部分と肉付部分を図３に示す。

```
U_SENT_1 ─┬─ PRED_1
          └─ (CASE_1 CASE_2 CASE_3)
PRED_1 ──── VERB_1
CASE_1 ──── S
CASE_2 ──── DO
CASE_3 ──── PP ──── U_PP_1
U_PP_1 ──── PREP_1

VERB_1 ──── spelling - "propose"
CASE_1 ──── instance - N1
CASE_2 ──── instance - N2
U_PP_1 ──── object ── N3
PREP_1 ──── spelling - "to"
```

図３　スケルトン－フレッシュ型構文意味辞書の骨格部分(上)と肉付部分(下)

Skeleton-Flesh Type Semantic Structure Dictionaries for Japanese-to-English Machine Translation

Akio YOKOO (Email:ayokoo@nttkb.ntt.jp), Hiromi NAKAIWA, Satoshi SHIRAI and Satoru IKEHARA
NTT Network Information Systems Laboratories, 1-2356 Take Yokosuka-shi, Kanagawa, 238-03 Japan

3．2　骨格構造のカバー率

3．1で示した考え方に基づき，既存の英語パターン辞書（結合価パターン，慣用表現辞書，ダ文述語パターン，計13,225）を再構築した。その結果，全体を446種類の骨格構造(ｽｹﾙﾄﾝ)で記述することができた。骨格構造の出現頻度の高い方から見た順位とカバー率（そこまでの骨格構造の累積で，全パターンのどれだけを表現できるかの百分率）の関係を表1に示す。

表1　骨格構造の出現順位とカバー率の関係

出現順位	9	17	40	95	214	314	446
ｶﾊﾞｰ率	70.6	80.7	90.0	95.0	98.0	99.0	100

骨格構造のうち出現頻度の高い10種類について，その骨格構造，出現頻度，頻度累計，カバー率を表2に示す。また，頻度順上位100までの骨格構造のカバー率を図4に示す。

表2　骨格構造(ｽｹﾙﾄﾝ)の上位１０構造
（全パターン数：１３２２５）

順位	骨格構造（ｽｹﾙﾄﾝ）	頻度	累計	ｶﾊﾞｰ率
1	Nx Vt Nx.	2622	2622	19.83
2	Nx Adj.	1893	4515	34.14
3	Nx Vt Nx prep Nx.	1436	5951	45.00
4	Nx Vi prep Nx.	1083	7034	53.19
5	Nx Adj prep Nx.	981	8015	60.60
6	Nx Vt noun prep Nx.	406	8421	63.67
7	Nx Vi.	402	8823	66.71
8	Nx Vt noun.	272	9095	68.77
9	Nx Vi prep Nx prep Nx.	240	9335	70.59
10	Nx be Vpp prep Nx.	231	9566	72.33

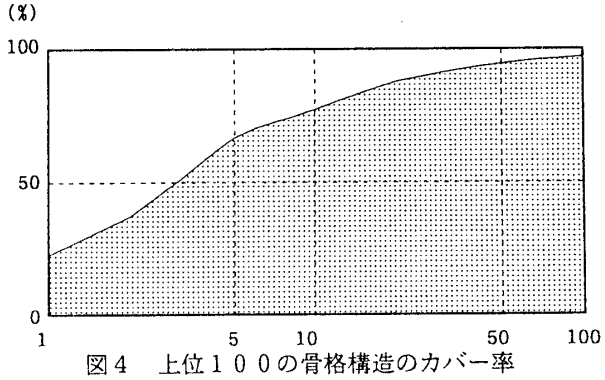

図4　上位１００の骨格構造のカバー率

以上から，少数の骨格構造により効率よく英語パターンを記述できるということがいえる。辞書メンテナンス・ツールでは，比較的出現頻度の高い52種類の骨格構造を「高頻度骨格構造」としており，これらで12,084パターン，91.37%をカバーしている。

4．利用者用構文意味辞書への応用

ALT-J/Eでは，結合価パターン辞書は，一般的な用言を格納した一般パターン辞書と，分野に依存した用言を格納した専門パターン辞書で構成される。専門パターン辞書は，ある意味では利用者が自分で作成する利用者パターン辞書と捉えることができる。そこで，専門パターン辞書（全体で206パターン）について，骨格構造の出現頻度とカバー率を調査した。その結果を図5に示す。

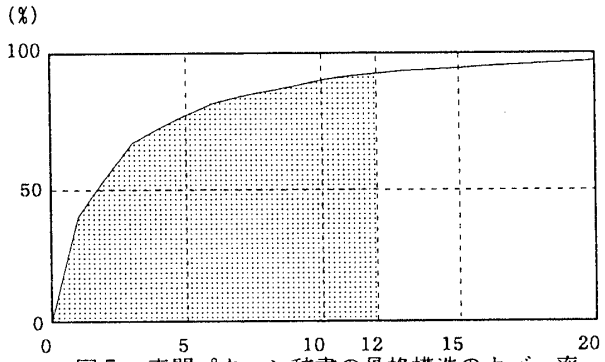

図5　専門パターン辞書の骨格構造のカバー率

上図のうち，頻度順12位までの骨格構造(カバー率：92.7%，網掛け部分)は，高頻度骨格構造に含まれており，利用者が個人用のパターンを作成するときには，既存の骨格構造を参照しながら，格要素の変数名や単語の字面だけを指定すればよいので，新規登録の効率化を図ることができた[3]。

5．おわりに

本方式により，既存の13,225のパターンを446の骨格部分に分類することができ，出現頻度が40位までの骨格構造で全体の90%をカバーできることが分かった。また，利用者が構文意味辞書を登録する場合もこの40の構造に入ることが多く，利用者辞書を構築する上での負担を軽減できることを示した。

また，日本語の単一用言文が英語の500文型程度で記述可能なことが分かったので，今後は，本記述形式による辞書の英文解析への応用についても検討する予定である。

〈参考文献〉

[1]白井，池原，横尾，中岩：「前編集不要の日英機械翻訳の実現に向けて」，NTT R&D, Vol.40, No.7, pp.897-904, 1991

[2]池原，宮崎，横尾：「日英機械翻訳のための意味解析用の知識とその分解能」，情処論，Vol.34, No.8, pp.1692-1704, 1993

[3]白井，横尾，池原，井上：「日英機械翻訳用構文意味辞書の記述精度の向上と作成支援」，48情処全大 6Q-9, 1994