日英機械翻訳における用言の訳し分けと構文意味辞書

白井　諭[†]　井上浩子[‡]　横尾昭男[†]　池原　悟[†]

[†]ＮＴＴコミュニケーション科学研究所　[‡]ＮＴＴアドバンステクノロジ

1 はじめに

意味解析技術の重要性が指摘されているが，機械翻訳では，動詞と名詞の意味的共起を結合価パタンとして記述し，それを原言語と目的言語で対にしたパタン対（構文意味辞書）の使用が有効であることが知られている。この方法には，パタン記述精度とパタン対収集の２つの課題がある。

パタン記述精度の問題については，日英機械翻訳では，名詞の意味属性を2,000種類以上に分類すれば，日本語の用言を訳し分けるのに必要なパタン対が記述できることが知られている[池原93]。

パタン対収集の問題についても，種々のヒューリスティックスや学習技術を応用した方法が提案されている。例えば，黒橋らは例文とシソーラスを用いて文型を同定する方法を提案し，精度は人手作成の方法と同等以上である可能性を示している[黒橋92]。また，アルモアリムらは学習技術を用いたパタン対の自動抽出方法を提案し，６動詞に対し各27～80の対訳用例を用いたパタン対の抽出実験に成功している[Almuallim 94a][Almuallim 94b]。しかし，日英翻訳では使用頻度の高い和語動詞のパタン対の機械学習だけで1,000万ペアの単純な日英対訳文が必要であると言われており[金田94]，作成すべきパタン対の総数が不明であること，パタン対の網羅的収集に必要な単純な用例を多数収集するのは困難であることなどの基本的な問題が未解決である。

これに対し，訓練されたアナリストであれば，類推により１例文から１パタンを作成することが期待される。そこで本稿では，人手による３つのパタン対作成方法を実験し，日英機械翻訳にはどれだけの数のパタン対が必要か，どのようにすれば作成できるかについて和語動詞1,000語を中心に検討する。そして，その結果を踏まえることにより，パタン対全体の必要数についての推定を試みる。

2 パタン対収集の方法

結合価パタンを記述するには，格要素となる名詞が2,000種類以上に分類されている必要がある[池原93]。本稿ではこの条件を満たしている日英翻訳システムＡＬＴ－Ｊ／Ｅ[池原89]の枠組みを対象にパタン対の作成方法を検討する。

2.1 和英辞書を参照する方法

(1) パタン対収集の方法

パタン対を収集する第１の方法として人間用の辞書の使用が考えられる。和英辞書には，日本語の語句に対し，語義ごとに対応する英語の語句，語法，例文などが記載されている。従って，用言の項目に対する語法や例文を分析し，格要素，副詞要素などの日本語側の制約条件を整理すれば，パタン対を作成することができる。

(2) 収集されたパタン対の数

数冊の和英辞書から得たパタン対は，当初，汎用パタン対10,000件，慣用パタン対5,000件であった[池原93]。その後，類似の汎用パタン対の統合や慣用パタン対の一部の汎用パタン対化により，汎用パタン対10,000件と慣用パタン対3,000件から成る構文意味辞書が完成した[横尾94]。

(3) 翻訳実験における充足性

この構文意味辞書を使用して，情報処理装置関連の仕様書（1,361文）の翻訳実験を行なったところ，試験文中に現れた用言の種類は142件で，辞書から使用されたパタン対は，120用言に対する154件であった。試験文中の22の用言（22パタン）はパタン対が登録されておらず，23の用言に対しては合計25のパタン対が不足していることがわかった[白井94a]。中でも，パタン対が不足している用言は，単語当たりの語義の多い和語動詞が多く見られた。

A Japanese to English semantic structure dictionary to select translation patterns for Japanese predicates

Satoshi SHIRAI[†], Hiroko INOUE[‡], Akio Yokoo[†] and Satoru IKEHARA[†]

[†] NTT Communication Science Laboratories and [‡]NTT Advanced Technology Corporation

2．2 日本語辞書の語義を参照する方法

(1) パタン対収集の方法

和語動詞は語義が多いため，パタン対を網羅的に収集するのは容易でない。和語動詞については，従来，日本の言語学者（20名あまり）を中心に，その語義と対応する用例を収集分析する研究が進められており，漢字表記の揺らぎを除く861動詞に対して，語義と語義毎の用例がＩＰＡＬ動詞辞書[IPA 87]としてまとめられている。

そこで，第２の方法としてこの辞書を使用したパタン対の収集を考える。具体的には，ＩＰＡＬ動詞辞書の各語義に示されている用例文に対し，日本語原文に忠実で，かつ英語として十分通用する英訳文を翻訳家に作成してもらい，その対訳から構文意味辞書に登録するパタン対の収集を試みる[白井94b]。

(2) 収集されたパタン対の数

上記の方法では，861の和語動詞に対して，5,243文（和文75,000字，英文40,000語）の例文が得られた。現在，パタン作成作業は，実施中であるが，これらのうちの740動詞に対する4,500例文から1,290パタン対が新規に作成され，既存のパタン対のうち，410件が修正された。逆に，第１の方法で得られたパタン対には，第２の方法では得られないものがかなり多く含まれていることもわかった。

これらの結果から，第１の方法，第２の方法のいずれの場合にも，得られたパタン対には欠落が多いことがわかる。

(3) 追加拡充の程度

ＩＰＡＬ用例辞書は，日本語動詞の語義分類に基づいて，用例が作成されている。従って，日英翻訳用のパタン対の観点からは，日本語動詞の語義とパタン対との対応関係（１語義が１パタンに対応するか）が問題となる。そこで，日本語の語義の多い代表的な４動詞について，語義とパタン対の対応関係を調査した（表１[白井94b]）。その結果によれば，両者が１対１に対応するものは，４割にとどまり，両者は必ずしも対応しないことがわかった。即ち，ＩＰＡＬ辞書の語義分類は，英語に訳出する上では必ずしも適切でないことを意味している。日英翻訳では，日本語と英語の意味的対応関係に即して，日本語を分類する必要のあることがわかる。

表１　ＩＰＡＬ語義とパタン対の対応

分類／動詞	「語義」対「パタン対」の関係					合計
	1対1	1対n	m対1	m対n	保留	
あがる	8	5	1	3	1	18
あげる	14	2	1	1	3	21
だす	8	9	5	4	1	27
でる	13	3	10	4	2	32
合計	43 43.9%	19 19.4%	17 17.3%	12 12.2%	7 7.1%	98 100%

2．3 人の知識を内省する方法

(1) パタン対収集の方法

前節までの結果から，人間用の和英辞書，日本語辞書の双方からパタン対を収集しても，まだ十分な数が得られないことがわかった。日本語例文とパタン対の関係を観察したところ，同じ動詞を使用していても，動詞の使われ方のニュアンスが異なるときに，新たな英語パタンが必要となる場合が多いことに気づく。そこで第３の方法として，英語の理解できる日本人が辞書等を手がかりにして自分の知識を引き出し，日本語としてニュアンスの異なる用法を可能な限り列挙するという方法で用例の収集を試みることにした。

列挙する用例は，作業にかける時間にもよるが，ある程度の時間以上考えても用例が思い浮かばなくなるまで抽出することとした。用例数の具体的な抽出目標数としては，いくつかの動詞に対する試行結果に基づき，ＩＰＡＬ動詞辞書の語義数の３～４倍に設定した。また，これらの日本語用例に対する英訳は，３．２節と同様に翻訳専門家に依頼して対訳用例集を作成することとした。

(2) 収集されたパタン対の数

上記の方法に基づけば，約半人年の作業により，300動詞（漢字表記の異なり見ると450動詞，語義数では1,700件）に対し，用例5,200文（和文65,000字，英文34,000語）が収集された。また，収集した用例のうちランダムに選択した30動詞（1,100用例文）からパタン対の抽出を試行したところ，新たに300パタンが抽出された。第１，第２の方法で得られなかったパタン対が，動詞当たり平均10パタンも見つかったことになる。

3 パタン対収集方法の評価

3.1 用例数とパタン対数の比較

適当に選んだ約30の和語動詞に対して２節で述べた３方法を適用し，得られた用例数とパタン対数を比較した（表２）。表から以下のことがわかる。

①第１の方法に加えて第２の方法を実施すれば，第１で得られるパタン対の数の２倍近くのパタン対が収集できる。

②第１，第２の方法に加えて第３の方法を実施すれば，第１，第２で得られるパタン対の数のさらに２倍以上のパタン対が収集できる。

以上から，和語動詞についていえば，第１の方法では1,500動詞に対する4,000パタン対が収集され[池原93]，第２の方法では1,000動詞に対し1,500パタン対が新たに収集され，第３の方法では1,200動詞に対する4,000パタン対がさらに収集される見込みである。

なお，第３の方法による用例収集では網羅性が問題となるが，一部の動詞について別のアナリストが用例を作成したところ，格要素の名詞などは異なるものの，パタン対を作成する上では等価とみなせる例文が再現率90%以上で収集されたので，アナリストによる差異はあまり問題ではないと思われる。

3.2 必要なパタン数と用例数の見積もり

日英機械翻訳において，パタン対記述の対象となる述語には，和語動詞のほか，サ変動詞や形容詞系の述語がある。サ変動詞の場合は１動詞あたり１～２パタンのものが多いと予想され，内省による用例文の作成は比較的容易であり，第３の方法によるパタン対収集が可能である。形容詞系の述語は和語動詞と同様に多義性があり，第２，第３の方法によりパタン対を収集する必要がある。

以上を踏まえてパタン対の数を推定すると表３を得る。表３では，推定されたパタン対の数に対し本稿の方法によりどれだけ収集できる見込みかも示す。この表から，日英機械翻訳では，汎用パタン，慣用パタンを含めると，約25,000のパタン対が必要と推定されること，また，これらのパタン対は，第１，第２の方法に加えて，第３の方法を実施することにより収集可能となることがわかる。

表２　収集されるパタン対の数（和語動詞の場合）

IPAL表記	漢字表記	方法1で収集	方法2で収集 語義数	方法2で収集 例文数	方法2で収集 追加P	方法3で収集 用例数	方法3で収集 新規P	一般Pの合計	慣用P
あく	空く	4	10	12	4	13	4	12	1
	開く	5			2	12	4	11	0
あける	空ける	4	11	17	1	14	5	10	0
	明ける	4			0	9	2	6	1
	開ける	3			1	9	2	6	1
あがる	上がる	7	18	31	16	90	16	39	12
あげる	上げる	8	21	31	13	98	16	37	14
いれる	入れる	5	19	30	12	113	28	45	10
うまる	埋まる	2	5	6	2	5	1	5	0
うずまる			3	4					
うめる	埋める	3	4	5	1	9	0	4	1
うずめる			4	4					
おちる	落ちる	8	11	21	7	53	23	38	1
おとす	落とす	6	14	19	5	53	15	26	3
きまる	決まる	3	8	17	2	32	10	15	2
きめる	決める	3	14	20	4	28	5	12	0
くずす	崩す	6	4	4	2	8	2	10	0
くずれる	崩れる	5	4	6	2	13	4	11	0
さく	割く	2	5	7	0	4	2	4	0
	裂く	1			5	3	0	6	0
さく	咲く	1	1	1	0	3	2	3	0
さける	避ける	3	6	11	0	9	2	5	0
さける	裂ける	1	1	3	2	4	1	4	0
たつ	立つ	5	13	24	4	75	30	39	11
	発つ	2			0	6	1	3	0
	建つ	1			0	5	0	1	0
	経つ	2			0	3	0	2	0
たつ	断つ	4	1		0	6	1	5	0
	絶つ	4	3		0	6	2	6	0
たてる	立てる	8	9	17	7	69	29	44	7
	建てる	1	1		0	5	0	1	0
だす	出す	16	27	53	15	95	22	53	21
でる	出る	22	32	49	5	145	38	65	18
はいる	入る	7	23	34	11	105	31	49	5
合　計		156	271	426	123	1102	298	577	108

4 おわりに

日英機械翻訳において，用言の意味を訳し分けるのに必要なパタン対の数とその収集手段について述べた。具体的には，単語当たりの語義が多くパタン対作成が困難な和語動詞を中心に検討し，①和英辞書からパタン対を収集する方法，②日本語動詞の語義対応の用例文を英訳してパタン対を抽出する方法，③それらを参考に，人の知識に基づいて作成した用例文を英訳してパタン対を抽出する方法，の３つのパタン対用例収集の方法を比較した。その結果，主要な約1,000の和語動詞を意味に応じて訳し分けるには，7,500件のパタン対が必要であることがわかった。また，従来の和英辞書から収集できるパタン対の数は必要数の約1/4であること，和英辞書と日本語辞書の語義分類を使用する場合は約1/2であること，必要なパタン対を網羅的に収集するには人の知識を内省して用例を作成する方法が適していることなどがわかった。

また，上記の結果から推定すると，サ変動詞，形容詞系の述語，用言性慣用表現などを含むパタン対全体では，約25,000パタンが必要であること，それらのパタンも人の知識を内省する方法で抽出された用例から比較的容易に収集できる見込みであることがわかった。

現在，第１の方法で得られたパタン対を拡充するため，第２，第３の方法を並行して実施中であり，和語動詞，サ変動詞，形容詞系述語に対してそれぞれ，5,000件，4,000件，2,000件（合計11,000件）のパタン対を収集済みである。また，慣用パタンとしては3,000件が収集済みである。今後は，残されたパタン対（汎用パタン約9,000件，慣用パタン2,000件）を整備していく予定である。

＜謝辞＞
本検討にご協力くださった山本弥生氏を始めとするＮＴＴアドバンステクノロジの各位に感謝する。

＜参考文献＞

[Almuallim 94a]Almuallim,H., Akiba,Y., Yamazaki,T., Yokoo,A. and Kaneda,S.: A Tool for the Acquisition of Japanese to English Machine Translation Rules using Inductive Learning Techniques, CAIA 94, San Antonio, Texas
[Almuallim 94b]Almuallim,H., Akiba,Y., Yamazaki,T. and Kaneda.S.,; Two Methods for Learning ALT- J/E Translation Rules from Examples and a Semantic Hierarchy, The Proc. of the 15th International Conference on Computational Linguistics, Kyoto
[池原89]Ikehara, S.: Multi-Level Machine Translation System, Future Computer Systems, Vol.2, No.3
[池原93] 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
[IPA 87] 情報処理振興事業協会 技術センター：計算機用日本語基本動詞辞書ＩＰＡＬ(Basic Verbs)，解説編＆辞書編
[金田94] 金田，秋葉，石井，アルモアリム：事例に基づく英語動詞選択ﾙｰﾙの修正型学習方式，「自然言語処理における学習」シンポジウム論文集，信学会＆ソフトウェア科学会
[研究社84] 小島，竹林編：ライトハウス和英辞典，第２版，研究社
[黒橋92] 黒橋，長尾：格フレーム選択における意味マーカと例文の有効性について，情処研報NL-91-11
[白井94a] 白井，横尾，池原，井上：日英翻訳用構文意味辞書の記述精度の向上と作成支援，第48回情処全大6Q-9
[白井94b] 白井，横尾，池原，井上：ＩＰＡＬ動詞辞書との対比による日英翻訳用構文意味辞書の充足性の検討，1994年秋期信学全大D-63
[横尾94] 横尾，中岩，白井，池原：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，第48回情処全大6Q-8

表３　日英機械翻訳で必要なパタン対の数とその収集に必要な用例の数の見積もり（＊ 漢字表記による異なり数）

比較項目 / パタンの種類		必要量の見込み		第１の方法(結果)		第２の方法(追加見込み=作業中)			第３の方法(追加見込み=一部実施)		
		用言数	パタン対の数	用言数	パタン対の数	*対象用言数	用例数	パタン対見込	対象用言数	見込み用例数	パタン対見込
汎用パタン対	和語動詞	1,500	9,500	1,500	4,000	1,000	5,200	1,500	1,200	15,000	4,000
	サ変動詞	6,500	8,000	3,000	4,000	---	---	---	4,000	8,000	4,000
	形容詞系	2,000	3,000	1,100	2,000	200	2,400	500	500	2,000	500
	小計	10,000	20,500	5,600	10,000	1,200	7,600	2,000	5,700	25,000	8,500
慣用パタン対		---	5,000	---	3,000	---	---	---	---	不明	不明
合　計		10,500	25,500	5,600	13,000	1,200	7,600	2,000	5,700	25,000	8,500