

「高度メディア社会の生活情報技術」
平成13年度採択研究代表者

池原 悟

(鳥取大学工学部 教授)

「セマンティック・タイポロジーによる言語の等価変換と生成技術」

1. 研究実施の概要

高度なネットワーク社会において言語バリアフリーの通信を実現することは、緊急かつ重要な問題となっている。しかし、従来の要素合成法を基本とする自然言語処理技術は、すでに技術的限界に近いレベルにあり、これを応用した機械翻訳の品質もほぼ限界と見られる。

本研究は、この限界を突破するため、新しい原理に基づく言語意味処理の基礎を確立しようとするものである。「人間の対象把握作用には、思考形式とも言うべきある種の枠組みが存在し、それが言語表現に反映される」と言うセマンティック・タイポロジー（意味類型論）の観点から、言語表現の構造と意味の関係を意味類型化された言語知識データベースとして体系化し、言語の等価変換と生成の技術を実現する。

現在までの成果は以下の通りである。まず、約30種類のドキュメントから100万件の日英対訳文を抽出し、その中から、述部2または3カ所を含む重文、複文、複重文を合わせて15万件抽出すると共に、文型パターン記述言語を設計した。それに基づいて、上記15万件の対訳例文を対象とする単語レベル、句レベル、節レベルの汎化作業を行い、「日英文型パターン辞書（文型数22万件）」を作成した。また、「文型照合プログラム」を試作して文型パターン辞書の被覆率特性を調査し、文法レベルでの文型パターン記述方法の問題点を分析した。その結果に基づき、「意味レベル文型パターン辞書（暫定版）」を試作した。

次に、意味類型化の作業としては、まず、重文複文を構成する節の意味と節間の関係の意味を分類するための2種類の意味分類体系（いずれも深さ5段、分類数約400種類）を完成した。引き続き、全文型パターンにこれらの意味分類コードを付与することにより、「日英意味類型パターン辞書（暫定版）」を作成した。また、これらの意味分類コードを使用した文型パターン意味検索プログラムを実現した。

今後は、「意味レベル文型パターン辞書」の被覆率の向上を図ると共に、「意味類型パターン辞書」に対する意味的検索精度の向上を目指して、意味分類体系の精密化を実現する。

2. 研究実施内容

今年度実施した主な研究項目は以下の通りである。

(1) 適合率向上を目指した「文型パターン辞書」の改良

昨年度までに開発した「文型パターン辞書（3レベル合計22万パターン対）」を対象に、文型パターンパーサーを使用した各種の被覆率評価実験を行い、適合率（被覆率）の向上を目指して様々な改良をおこなった。最大の改良項目は、述部語尾の「時制・相・様相」情報の汎化である。これらの汎化では、関数記述に関する従来の文型パターン記述の仕組みを大幅に拡張した。新たに導入した仕組みの主なものは「関数の任意化と要素選択記号化の仕組み」、「字面要素への関数付与の仕組み」、「対応型任意記号と対応型要素選択記号の導入」、「時制関数の導入」、「グループ関数の導入（＜同値型グループ関数＞、＜対応型グループ関数＞）」である。このほか、被覆率向上の対策として実施した主な項目を挙げると、「自立語字面の異表記の吸収」、「各種修飾要素の任意化」、「言い換え可能表現の選択記号化」、「否定表現の任意化」、「丁寧表現の任意化」、「限定詞変数の適用」などである。

(2) 意味的正解率向上を目指した「文型パターン辞書」の改良

意味レベル文型パターン辞書では、昨年度、名詞の意味的制約条件として「日本語語彙大系」の名詞意味属性体系（2700種）を使用し、用言の意味属性体系としてALT-J/Eの用言意味属性（36種）と「類語辞典」（講談社）の意味分類（1000種）を使用した。しかし、文型パターンの意味的絞り込みの実験においてこれらの意味分類体系では、期待するだけの効果が得られないことが分かった。

そこで、問題点を分析し、文法的に適切な日本語文型パターンに適合させることを目的とした「文法的役割」と意味的に適切な英語文型パターンを持つ日本語文型パターンに適合させることを目的とした「意味的役割」の双方を両立させる立場から、重文・複文の文型パターン専用、用言と体言に対する新しい意味分類体系を設計し、用言意味辞書と体言意味辞書を作成した。また、それらの意味辞書を使用して、文型パターンの名詞変数と用言変数に対して、新しい意味コードを付与した。これらの効果は、新年度実験評価する予定である。

また、意味的正解率の問題分析では、英語文型記述方式と離散記号の指定方法に様々な問題が残されていることが分かったため、その改良を行った。まず、英語文型記述方式では、主として、①前置詞の記述法、②所有格代名詞の記述法、③省略された目的語の記述法の改良を行った。また、離散記号付与基準については、意味的に不適切な文型パターンへの適合を誘発している離散記号の種類と付与場所を分析し、その書き換えを行った。

(3) 「日英基本文型パターン辞書」の検討

日英文型パターン辞書から主要な文型パターンを抽出し、①ブラッシュアップ作業、②類似文型パターンの縮退と制約条件の見直し修正作業、③基本文型パターンの体系化作業を行い、基本文型パターン辞書（暫定版）を作成した。

まず、①では、文型パターン作成基準に従って作成されているかを確認した。作成基準に適合しないものは修正した。②では、より以上に汎化できないかどうかを調べ、汎化可能な部分を汎化すると共に、他の文型パターンに包含されるような不適切な文型パターンを捨てた。②では、上記で得られた文型パターンファイルを使用して、再度、全標本文を対象とした3種類のクロスヴァリデーションテストを行い、「頻度順文型パターンリスト」3種類を作成し直した。新規に作成した「頻度順文型パターンリスト」の上位2000件について意味的な分類を行い、日本語基本文型パターンを取り出した。また、取り出した日本語基本パターンとそれに対応する英語パターンを調べ、1対nの関係を持つ文型パターンを表に纏めた。

(4) 「文型パターン意味類型化辞書」の試作

まず、重文・複文を意味分類するため以下の2種類の意味分類体系を作成した。

①単文の意味分類体系

述部の用言（動詞、形容詞、形容動詞）に着目して、命題レベルの単文（節）を意味分類するための体系を新規に作成した。分類数は370種類、分類の深さは5段とした。また、この分類体系に基づく「日本語用言意味辞書」を作成した。収録する対象は、日常的に使用される一般語をほぼ網羅する範囲（6,000語）である。なお、用言意味辞書は、多義を考慮した構成とした。

②重文複文の意味分類体系

田窪・増岡の従属節分類（90種）を重文複文の意味分類の観点から見直し、「重文複文意味分類体系」を作成した。分類数は150種類、分類の深さは5段とした。

次に、文型パターン辞書のすべての文型パターンに上記の意味分類コードを含む以下のような文法的、意味的な分類コードを付与した。

「文種別コード」、「統語構造第1分類コード」、「統語構造第2分類コード」、
「第1従属節意味コード」、「第2従属節意味分類コード」、「節の意味分類コード」、
「第1節間意味分類コード」、「第2節間意味分類コード」、「第1節間KWコード」、
「第2節間KWコード」

(5) 「文型パターンパーサー」の大幅改造

すでに、試作した「文型パターン照合プログラム」の経験を元に、「文型パターンパーサー」を試作したが、その後、様相・時制などの述部語尾表現の汎化を行ったことなどにより文型パターン記述言語仕様の大幅な改造が必要となった。そこで、それに合わせて、文型パターンパーサーの改造を行った。改造・改良した項目は主として以下の5種類である。

①パターン記述言語仕様およびパーサの強化

文型パターン辞書の「時制表現」、「様相表現」について柔軟にパターンを照合するために、文型パターン記述言語の言語仕様を拡張しパーサを強化した。

②定義パラメータの強化

パターン記述言語に用いる関数・変数・記号の言語構造の定義は、定義ファイルとなっており、パーサの処理条件となる。「様相時制」の関数定義も定義ファイルにより与えるので、記述能力を強化するため、定義ファイルの記述仕様を見直した。

③意味属性体系の拡張性の強化

意味属性体系として、2004年度前期には、「日本語語彙大系」および「講談社類語大辞典」に基づく体系を用いていた。これに対して、「新たな意味属性体系」を作成し、照合実験を行うための意味属性の扱いに関する処理系の整備を行った。

④システム構成の修正

照合実験の効率化および実験環境の範囲拡大を目指して、システム構成を修正した。具体的には、ALT-J/Eのコードを用いた解析処理が含まれていたようなコードを除去した。また、定義パラメータの変更のたびに照合用ファイルの再生成が必要でありhour オーダーで待ち時間が発生していたが、定義パラメータを変えた照合実験を何度も行うので、各種の照合用ファイルの関係について修正した。

(6) 「文型パターン意味検索プログラム」の試作

前項の文型パターンパーサーは、入力文と文型パターン辞書を照合し、入力文の構文構造と表現要素の文法的、意味的制約条件に一致した文型パターンをすべて抽出するプログラムである。これに対して、ここでは、入力文と意味的に類似と見られる文型パターンを検索するプログラムを試作した。前者は形式的に一致する文型パターンのみが検索されるが、後者は、形式は異なっても意味的に類似と判定された文型パターンが検索される。

本研究では、「真理項」を使用した文型パターン写像方式の実現を目指しているが、このプログラムは、この方式の実験的検討を行うためのものである。すなわち、このプログラムで、検索に使用する意味分類コードは一種の「真理項」と考えて良いから、このプログラムは、日英文型パターンの写像関数の暫定版と見なすことができる。どれだけの品質の意味検索ができるか不明であるが、このプログラムを使用した検索実験によって、将来の写像関数の設計する際の重要な情報が得られると期待される。

なお、このプログラムは、人手翻訳の支援にも使用できると期待される。すなわち、文型パターン照合プログラムによる文型パターン検索では、構成条件を厳密に満足する文型パターンのみが検索されるから、文型パターン辞書の中に入力文の翻訳に使用できる文型パターンや参考になる文型パターンが存在しても、それらが必ず抽出されるとは限らない。そのような場合にこのプログラムを使用すれば、入力文と意味的に類似の文型パターンが検索できる。

(7) その他の検討

その他、文型パターン辞書に直接反映する段階にない項目として様々な検討を実施した。その主なものを以下に示す。

①最適照合方式の検討

意味レベル文型パターンにおいて、文型パターンは意味的排他性が保証されることを条

件に、最も緩い制約条件で記述されることが望まれる。これに対して、現在は、変数の意味的制約条件として、標本文の単語の意味属性を付与している。すなわち、付与した意味属性は各意味属性体系の持つ意味属性の中で最も粒度の細かいものである。このため、この制約条件下での文型照合では、意味的に正しい文型パターンが漏れてしまう危険性がある。そこで、今年度初め、「最適照合方式の検討」として、最適な制約条件の粒度を発見するための文型照合実験を行ったが、現状では単語レベルの文型パターンの「文型再現率（R1）」が低く、時制などの汎化を優先して進めることが必要と判断されるため、この検討は一時、中断し、多変量解析を応用した最適文型パターン選択方式の検討を行った。その結果、現状の情報でもかなりの精度で意味的に適切な文型パターンが選択できるめどを得た。

②日英文型パターン間の意味的なリンク張りの検討

日本語1文型パターンに対して、1つ以上の英語文型パターンを対応づけることは、CREST研究の課題の一つでもある。現段階では、文型パターンの意味類型化は完成していないので、同一意味グループ間としてこの関係を決定することはできない。そこで、日本語文型パターンのひとつひとつに対して、意味的に同等な英語パターンを発見し、リンクを張る方法として、「元の日英対応関係を直接利用する方法」、「日本語入力文との照合実験による方法」、「文型パターンの意味分類コードを利用する方法」の3つの方法を検討した。

③文型パターンの縮退化実験

文型照合実験の結果を見ると、形式的、意味的に重複するような文型パターンが数多く含まれている可能性がある。どの程度圧縮できるか分からないが、将来、書籍として編集することを考え半分程度に縮退させたいものである。そこで、日英文型パターン間の縮退化の方法について検討し、包含関係にある日本語文型パターンを対象に、下記の手順による縮退化プログラムを作成した。

- ・対応する英語文型パターンが「一致関係」にあれば縮退させる。
- ・それ以外で、英語文型意味適合域が同じ英語文型パターンがあれば、縮退化させる。

しかし、現実の文型パターンで包含関係にあるものは予想以上に少なく（10%以下）、縮退化は容易でないことが分かった。このことから、現状ではまだ汎化不足が予想される。

④日英機械翻訳簡易実験システムの試作

文型パターン辞書の問題点を把握するため、変数、関数などの部分に対する英語を生成する機能からなる機械翻訳のための簡易実験システムの試作を開始した。現在までに結合価パターン辞書を使用した節変数の翻訳ルーチン、名詞句パターン辞書を使用した名詞句パターン翻訳ルーチンを作成したところである。

⑤シテ型接続文型の分析

単語レベルパターンから、文末に動詞シテ形接続を含むものを抽出し、見出し語で313について、名詞変数化のみのパターン約1600を作成した。各パターンには、約30の意味分類コードを策定して付与した。本パターン作成により極めて多義性の大きいシテ

形接続について、意味分類が可能となった。

3. 研究実施体制

研究統括Gグループ

- ① 研究分担グループ長：池原 悟（鳥取大学工学部、教授）
- ② 研究項目：基本方式の検討

等価変換システム研究グループ

- ① 研究分担グループ長：宮崎 正弘（新潟大学工学部、教授）
- ② 研究項目：言語等価変換システムの試作実験

言語生成システム研究グループ

- ① 研究分担グループ長：奥村 学（東京工業大学大学院精密工学研究所、助教授）
- ② 研究項目：言語生成システムの試作実験

意味類型知識ベース研究グループ

- ① 研究分担グループ長：池田 尚志（岐阜大学工学部、教授）
- ② 研究項目：意味類型知識ベース開発の開発

4. 主な研究成果の発表

(1) 論文発表

- 金出地真人、徳久雅人、村上仁一、池原悟：結合価文法による動詞と名詞の訳し分け能力の評価、自然言語処理、Vol. 11, No. 3, pp. 149-164 (2004)
- 池原悟、阿部さつき、徳久雅人、村上仁一：非線形な表現構造に着目した日英文型パターン化、自然言語処理、Vol. 11, No. 3, pp. 70-95 (2004)
- 池原悟、徳久雅人、竹内（村本）奈央、村上仁一：日本語重文・複文を対象とした文法レベル文型パターンの被覆率特性、自然言語処理、Vol. 11, No. 4, pp. 147-178 (2004)
- Masato Tokuhisa, Tsutomu Tanaka, Satoru Ikehara and Jin'ichi Murakami: Emotion Reasoning Based on Valency Patterns - a prototype annotation of causal relationships -, International Series on Natural and Artificial Intelligence, Vol.1, Human Intelligence Systems, From Control to Autonomy, pp. 534-539 (2004)
- 藤田篤, 乾健太郎, 松本裕治: 自動生成された言い換え文における不適格な動詞格構造の検出、Detection of Incorrect Case Assignments in Automatically Generated Paraphrases、情報処理学会論文誌, Vol. 45, No. 4, pp. 1176-1187, 2004.04
- Atsushi Fujita, Kentaro Furihata, Kentaro Inui, Yuji Matsumoto, and Koichi Takeuchi : Paraphrasing of Japanese Light-verb Constructions Based on Lexical Conceptual Structure. Proceedings of the 42th Annual Meeting of the

Association for Computational Linguistics (ACL) Workshop on Multiword Expressions: Integrating Processing, pp. 9-16, Jul., 2004.

- 竹信伸介、徳久雅人、村上仁一、池原悟：動詞節に修飾された名詞句の係り受け解析、情報処理学会、自然言語処理研究会、2004-NL-163、pp.83-89 (2004.9)
- Katsuyuki Shibata, Dependence on Context in case of English-Japanese Machine Translation I-1, Fukuoka University Science Reports, vol.35 No.1, pp.75-79, March (2005)
- Yoshihiko NITTA, Extracting Semantic Role Sequence from Bilingual Corpora by Shallow Parsing, Nihon University Economic Review, Vol.73, No.4, Nihon University(2004-1)pp.1-12
- 新田義彦, 外国語教育に自然言語処理を応用する方法の検討 (Applying Natural Language Processing Technique to Foreign Language Education) , Nihon University Economic Review, Yoshihiko NITTA, Extracting Semantic Role Sequence from Bilingual Corpora by Shallow Parsing, Nihon University Economic Review, Vol.74, No.1, Nihon University(2004-4)pp.1-12
- 新田義彦：参照構造ネットワークによるアイデア・プロセッサの方式検討 (Idea Processing Method by Referential Network Structure) , Nihon University Economic Review, Vol.74, No.2, Nihon University(2004-7)pp.1-17
- 新田義彦, 自然言語処理研究の潮流概観 (A Survey of Recent Trend of Natural Language Processing Study) , Nihon University Economic Review, Nihon University Economic Review, Vol.74, No.3, Nihon University(2004-10)pp.79-90
- Yoshihiko NITTA, Applying Natural Language Processing technique to Language E-Learning, Nihon University Economic Review, Vol.74, No.4, Nihon University(2005-1)pp.1-16
- 新田義彦, Web学習における教材提示方法の検討 (How to Manipulate Teaching Materials in Web-Learning Environment) , Nihon University Economic Review, Vol.75, No.1, Nihon University(2005-4)pp.1-20
- Yoshihiko Nitta, Masashi Saraki & Satoru Ikehara, Natural Language Approach to Language E-Learning, in: K.Murase et al (eds.), Human and Artificial Intelligence, Vol.1 , Fukui University (2004, Dec.5-6)pp.522-527