「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（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カ所を含む重文、複文、複重文を合わせて１５万件抽出した。引き続き、文法レベルでの文型パターン記述方式を検討し、文型パターン記述言語の言語仕様を定めた。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、句レベルの汎化作業を行い、それぞれ１２万件、10万件の日英文型パターンを作成した。

これらと並行して、因果関係を中心とした500例文を対象とした例題検討や因果関係以外の例文を含む250例文の意味記述などを行い、様々な観点から文型パターンの意味類型化の方法を検討した。また、入力日本文と文型パターンの包含関係を調べるアルゴリズの検討を行い、文型照合プログラム（第１版）を試作した。なお、このほかにも、究極的な機械翻訳の原理に迫るための言語表現と概念の関係に関する検討のほか、実験用もしくは、文型パターン補充用として、日本語基本文型の例文1万件に英訳を付与するなどの準備などを行った。

今後は、作成した文型パターン被覆率を調べ、被覆率向上のための各種対策を実施すると共に、文型パターンの意味類型化の方法論を検討し、実際に１５万文型の意味類型化を行う。また、日英変換実験システムと日本文生成実験システムを試作し、意味類型パターンの効果を実験的に検証する。

2. 研究実施内容

平成１４年度に実施した研究等の内容は以下のとおりである。

（１）対訳例文ファイルの作成

昨年度に引き続き、日英対訳例文55万件を収集し、コード変換などを行った後、出典情報などを付与し、標本データベースを作成した。また、そのデータベースに対して、文種別（単文、複文、複重文、疑問文、会話文、その他）に分類マークを付与したのち、重文、複文、複重文の３種類について、文字コードのチェックなどを行い、本研究で文型抽出用の標本セット（約10万件）を作成した。

（２）日英対訳コーパスの拡充

日本語文型の網羅性を向上させるため、「日本語文型辞典」（くろしお出版1998）の中から、日本語特有の表現を有する重文、複文１万件を抽出し、英語に翻訳した上で対訳標本ファイルを作成した。

（３）日英文型パターン記述方式の検討

文法レベルでの日英文型パターンを記述する方法について検討し、単語レベル、句レベル、節レベルの３段階からなる日英文型パターン記述言語を設計した。また、パターン記述言語に従った日英文型パターンを対訳例文を汎化することにより生成する方法について検討し、文型パターン作成手順書を作成した。作成した文型パターン記述言語は、様々な変数と関数を持つ言語で、汎用的な文型パターンの記述能力を持つ点で、従来のパターン記述方法に比べてかなりユニークで斬新なものと言える。また、日英対訳例文からの文型パターンの生成では、形態素解析結果や日英対訳辞書検索を応用した半自動的な汎化方法のめどを得ることができた。

（４）単語レベルと句レベルの文型パターンの作成

上記（１）で作成した文型パターン作成用標本文（重文、複文15万件）を対象に、（２）で作成したパターン記述言語の仕様に従って、単語レベルと句レベルの文型パターンそれぞれ13万件、10万件を試作した。本年度試作した文型パターン数の詳細は表３の通りである。

表1. 各レベルで作成された文型パターン数

	種 別	元のパターン数 （又は例文数）	作成パターン数 （又は例文数）	異りパターン数 （又は例文数）	備　考
－	対訳原文	15.5万件	15.7万件(*)	12.9万件(**)	*英文加工での増分2000件 **作業対象外は、2.8万件
1	単語レベル	12.9万件	129万件(*)	12.8万件(*)	*字面パターン600件を含む
2	句レベル	12.9万件(*)	10.5万件	9.9万件	*2.4万は句変数化対象外
3	節レベル ＜注＞	1.7万件(*) (10.5万件)	0.19万件 (1.2万件？)	0.16万件 (0.5万件？)	*1.5万は節変数化対象外 ()は、当面の予想値
－	合計	（３６万件）	（２５万件）	（２３万件）	

＜注＞節レベルの汎化を行ったのは、一部にとどまる。その結果から見ると、節変数化されるのは、句レベルパターンの11%程度であるので、絶変数パターンの総数は、1.2万パターン程度と見られる。また、現状の0.19万件の重複率が16%であることから大胆に推定すると、1.2万件に含まれる異なりパターン数は、5,000～6,000件定度と予想される。

（５）意味類型化方式に関する例題検討

意味類型パターン化の方法論を確立するため、昨年度作成した対訳例文ファイルの中の因果関係対訳例文（約500文）を対象に以下の作業を実施し、パターン化の問題点と方法論をまとめた。

① 文法属性の付与、② 名詞意味属性と動詞意味属性の付与、③ 文型の縮退、真理項概念ＫＷの付与、④ 日英文型のデフォールト対応付け、⑤ 日英文型マッピング情報の付与

また、因果関係以外の例文を含む250文を対象に、意味的論理範疇全体の構成法について検討を行い、真理項の種類と体系の構成案をまとめた。

（６）文型パターン照合プログラムの１次試作

入力日本文に対して、表現パターン辞書から、該当するパターンをすべて同時に抽出するアルゴリズムについて検討した。追加予算が得られたので、この検討結果に基づき、日本語入力文に対する文型パターン照合プログラムを試作した。

（７）文型パターンの被覆率推定方法に関する検討

例文からの文型パターン記述方法を決定するには、文型パターンの排他性を保ちながら、いかにして被覆率を向上させるかが極めて重要であるが、一般に被覆率は。文型パターンすべてが完成しないと測定することはできない。そこで、パターン収集の段階で最終的な被覆率を推定する方法について検討した。

（８）意味類型に基づく言語変換方式概念の検討

言語変換の基本原理を明らかにするため、概念論の立場から話者の認識と言語表現の関係について検討を行うと共に、言語間の概念を介した対応付けの原理について検討した。

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：言語生成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（論文発表および特許出願）

（１）論文（原著論文）発表

○ 池原悟（鳥取大学）：究極の翻訳方式の実現に向けて＝＝類推思考の原理に基づく翻訳方式＝＝、AAMT Journal，アジア太平洋機械翻訳協会、No.33，pp.1-7（2002.5）

○ 池原悟（鳥取大学）：「類推思考の原理に基づく言語の意味的等価変換方式」、鳥取大学総合情報処理総合センター広報、Vol.3、pp.13-35(2003.3)

○ 池原悟，村上仁一，宮本健司：「ＡのＢ」型名詞句の日英翻訳規則について，情報処理学会論文誌，Vol.43，No.7，pp.2300-2308（2002.7）

○ 池原悟、村上仁一、的場和幸：日英機械翻訳のための時間表現の意味と対応関係の解析、情報処理学会論文誌，Vol.44，No.2，pp.451-465（2002.2）

（２）特許出願

なし