研究課題別事後評価結果

1. 研究課題名： 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

2. 研究代表者名及び主たる研究参加者名（研究機関名・職名は研究参加期間終了時点）

研究代表者

池原 悟 （鳥取大学工学部情報システム工学科 教授）

主たる共同研究者

宮崎 正弘 （新潟大学工学部情報工学科 教授 （平成 13 年 12 月～））

奥村 学 （東京工業大学大学院精密工学研究所 助教授 （平成 13 年 12 月～））

池田 尚志 （岐阜大学工学部応用情報学科 教授 （平成 13 年 12 月～））

3. 研究内容及び成果：

従来の自然言語処理では言語表現の線形性を仮定した要素合成法が基本となっているが、現実には非線形な表現が多く、表現を単語に分解する過程で全体の意味が失われることが問題であった。この問題に挑戦するため、本研究では「意味類型論（セマンティック・タイポロジ）」と「等価的類推思考の原理」の2つの観点から言語の「意味的等価変換方式」を提案し、その実現に向けた研究開発を進めてきた。「意味的等価変換方式」は表現構造の持つ意味に着目し、与えられた言語表現を意味的に等価な別の表現に変換するもので、以下の2つのステップから構成される。

第1は「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フレームワークが存在し、それが言語表現に反映される」とする「意味類型論」（有田潤1987）の考えに基づくもので、意味のまとまる表現構造をパターン化して意味的に分類する。

第2は「人間の独創性は何らかの共通点を背景とした類推思考から生まれる」とする「等価的類推思考の原理」（市川亀久弥1963）を言語表現に適用するもので、与えられた表現構造を言語共軛な概念（真理項と呼ぶ）を介して他の表現構造に写像する。異なる言語間の表現に変換する場合は機械翻訳の技術となり、同じ言語内で変換する場合は「言い換え技術」となる。

以上の観点から、本研究では重文複文を対象とする「意味類型パターン辞書」を研究開発することを主たる目標とした。まずは「文型パターン辞書」を試作した後に、それを意味的に類型化することによって「意味類型パターン辞書」を開発した。その過程でそれらの品質を評価し、さまざまな改良を実施するために「パターン検索プログラム」を試作した。

（1）文型パターン辞書

言語表現とそれを構成する要素の線形性と非線形性の定義を明確にし、「すべての言語表現は0個以上の線形要素と非線形要素から構成される」とする非線形言語モデルを提案した。このモデルに基づいて「字面」、「変数（17種類）」、「関数（10種類151関数）」、「各種特殊記号（10種類）」の4種類の要素を用いて文型パターンを記述する文型パターン記述言語を設計した。また、重文複文のパターン記述の観点から約6千語の用言意味辞書と約6万語の名詞意味辞書を作成した。更に汎化対象となる線形要素の半自動的な判定方法を検討し、文型パターン作成手順の半自動化を図った。

これらの枠組みを用いて文型パターン辞書を作成するために、約30種類のドキュメントから比較的標準的な日英対訳例文約100万件を収集した。その中から述部2つ及び3つの重文と複文を抽出して形態素解析を行い、解析誤りを人手で修正の上、タグ付きの対訳コーパス(15万件)を作成した。次に、この対訳コーパスを対象に半

自動的な汎化手順を適用して、単語レベル（12.2万件）、句レベル（8.0万件）、節レベル（2．6万件）の文型パターン辞書（合計22．7万件）を作成した。

(2) 文型パターンの意味類型化

文型パターン辞書の意味類型化を行うため、重文複文の統語的構造に関する分類体系と意味の分類体系を構築した。また、すべての文型パターンに統語的分類コードと意味分類コードを付与した。

意味分類コードは文型パターンを意味類型化（意味的なグループ化）するために必須のものである。第1の分類体系として重文複文全体を構成する複数の節の意味的関係に着目した「節間意味分類体系（4段227種類）」を、第2の分類として個々の節の意味を表す「節の意味分類体系（5段742種類）」を開発した。

(3) パターン検索プログラム

意味類型パターン辞書の被覆率を評価し、問題分析と改良に役立てるため、「パターンパーサ」と「意味検索プログラム」を作成した。「パターンパーサ」は入力日本文と各パターンの構成要素を比較して、適合するパターンを発見するものである。15万件の標本文と22．7万パターン間の照合実験（クロス照合実験）を約30時間（入力文1文当たり1秒以下）で実行できる。「意味検索プログラム」は入力文の意味コードを判定し、それと同じ（又はその配下の）意味コードを持つ文型パターンを検索するものである。前者は入力文と一致する構造のパターンのみが検索されるが、後者は構造の異なるパターンも抽出されるため、日本語書き換えなどへも応用できる。

本研究の最終的な成果物である意味類型パターン辞書は実験の結果、統語的被覆率98.5%、意味的被覆率79.5%と実用的な水準を達成することができた。

本方式は意味的に非線形な表現構造を分解不能な単位とすることによって全体を線形近似に持ち込むものであり、意味処理の基本技術として多くの応用が期待される。

開発の過程で作成された日英対訳コーパス（100万文）はもちろんのこと、英語構文体系（83分類）、重文複文と英語構文の意味的対応表などは従来にない知的な資産であり、また、意味属性体系や単語意味辞書などは様々な意味処理に使用可能で、今後の多くの研究に役立つと期待される。

4．事後評価結果

4－1．外部発表(論文、口頭発表等)、特許、研究を通じての新たな知見の取得等の研究成果の状況

これまでにない新しいアプローチによる言語の意味処理に関する研究であり、日本語の重文複文の表現をほぼ網羅できるような意味辞書を実現した。100万件に及ぶ例文を収集すると共に、解析手法として非線形言語モデルを提案、このモデルに基づく文型パターン記述言語の設計、半自動的な判定方法の開発を行った。最終的に23万パターンに及ぶ意味類型パターン辞書を完成したが、これは今後の言語処理、機械翻訳に役立つ資産となるものである。この辞書を用いて日本語解析を行ったところ、予期以上のカバー率（約80%）を達成した。膨大な実データに基づく実証的な研究手法である。言語の意味処理技術として他に例のない研究であり、新しい道を切り拓いた先駆けとなる研究として高く評価したい。

計算およびメモリー資源が潤沢に得られるという技術的背景の変化によって、このようなアプローチの有用性を示し、言語処理の分野に新しいパラダイムを導入すると共に、その有効性について予備的な実証を行ったもので、意訳を含むより精度の高い機械翻訳への展開が期待される。また、日本語処理、自然言語処理の研究の基盤となる重要なデータの蓄積であり、この分野の今後の研究の進展に大いに役に立つと期待される。日本語処理に一つのマイルストーンを築いたといえるだろう。

論文発表は国内22件、国際5件、口頭発表は国内127件、国際19件と積極的に発表しているが、何れも工学系の学会である。対象が言語処理であることから、これからはいわゆる「文系」の言語学研究者にも理解してもらう努力が重要である。この成果を他の研究者にも判り易い表現で説明して理解を広めること、蓄えられた言語

資産をさらなる研究の展開のために公開することを要望したい。

特許出願はないが、本研究は工業所有権にはなじまない内容であることからやむを得ないと考える。しかしながら、研究によって得られた知的財産は大きく、今後これが十分に活用されることを期待する。

4－2. 成果の戦略目標・科学技術への貢献

本研究の成果はいずれも他に例が無く、この研究によって初めて得られたものであり、今後この分野の研究の出発点となる成果といえる。その意味での今後の自然言語処理研究に対する貢献は大きい。日英対訳文型パターン辞書や意味記述言語などの成果を適切に普及させれば、これが世界的な多言語応用にも広がっていく可能性を秘めている。そのためにも意味類型パターン辞書を含む本研究の成果をなるべく早期に公開し、その内容をよく広報して関係者の理解を深める努力を望む。

当面の実際的な応用としては精度の高い機械翻訳であるが、将来的には「計算機の言葉を人間が理解しなければならない社会」から「計算機が人間の言葉を理解する社会」への突破口となることが期待される。

4－3. その他の特記事項（受賞歴など）

人工知能学会業績賞（平成18年）など国内4件の受賞