「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。これと併行して、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、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、試作した「文型照合プログラム」によっ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。

今後は、被覆率調査実験によって「意味レベル文型パターン辞書」の問題点を明らかにし改良を進めると共に、その意味類型化の作業を開始する。

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）「文法レベル文型パターン辞書」（第１班）の試作

昨年度までに開発した単語レベルと句レベルの文型パターンに節レベルの文型パターンを加え、「文法レベル文型パターン辞書」（第１班）を作成した。15万件の対訳法本に対して作成された節レベルの文型はパターンは、1.3万件であった。標本文当たり２

～３個の節表現が含まれていることから見て、節表現の90%以上は、非線形であることが分かる。これは、節毎の翻訳結果を合成するような従来の要素合成法の方式では、対訳標本にあるような品質の良い英語翻訳は得られないことを意味しており、本方式の有効性を期待させるものとして重要な結果であった。

（２）「文法レベル文型パターン辞書」の被覆率実験調査

「文法レベル文型パターン辞書」（第１班）の被覆率特性を評価した。その結果によれば、任意の入力文の７割が単語レベルの文型パターンに適合し，９割が句レベルの文型パターンに適合するのに対して，節レベルの文型パターン全体の被覆率は低いことが分かった。節レベルの文型パターンは，被覆率が高くなることを狙って作成されたものであるが，実際の対訳例文で節レベルまで汎化できるものは少なかったためである．

また、入力日本文に対して，完全一致文型が存在する割合と部分一致文型しか存在しない割合を調べたところ、単語レベルに比べて句レベルの文型パターンでは、完全一致率が高いことが分かった。完全に一致する文型は意味的にも適切である可能性が高いと予想されることから、句レベル文型パターンの重要性が確認されたといえる。これに対して、節レベルの文型パターンの役割はあまり大きくなさそうである。

ところで、入力文に適合する文型パターンにおいて、適合の仕方は必ずしも一通りとは限らない。そこで、適合文型パターンが存在した入力文の場合について、どれだけの数の文型パターンが適合したかを調べると、句レベルで適合した異なり文型パターン数は、単語レベルの場合の約５倍で、文型パターン当たりの解釈数も２倍以上であること、また、節レベルの文型パターンの場合、異なり文型パターン数は比較的少ないが、逆に、文型パターン当たりの解釈数は多いことがわかった。

文法レベルの文型パターンでは、適合文型パターン数が多いことから、今後，意味的に精密な制約など，文型パターンに対して強力な適合条件を付与する必要のあることがわかる．

（３）「意味レベル文型パターン辞書」の開発

上記の実験結果に基づき、意味レベルの文型パターン記述方法を決め、「文法レベル文型パターンファイル」に収録された文型パターンの単語変数、句変数、節変数に意味的な制約条件として、意味属性を付与すると共に、意味的な汎化として、出現順序が任意な要素の指定、(2)位置変更可能な要素の指定、(3)文型パターン内に挿入しても良い要素の種類とその位置の指定、(4)その他表記の揺らぎの吸収などを行い、「意味レベル文型パターン辞書（暫定版）」を作成した。

文法レベル文型パターンに比べて、意味レベル文型パターンの再現率は、大幅に低下することを予想していたが、実験では、10-20%程度の差しかないようである。また、適合した意味レベルの文型パターンの意味的正解率はまだまだ期待通りの値は得られていないことが分かった。単語レベル、句レベル、節レベルの３レベルの文型パターンを組み合わせて使用すれば、50%程度の入力文は意味的にカバーされるため、とりあえずのカバー率としてはまずまずであるが、この結果は、まだ暫定的なものである。文型パ

ターンの改良も作業中であり、様々な汎化を実施している。また、文型照合方式についても、名詞意味属性や用言意味属性の詳細を指定する実験など、様々なケースについての実験を行って、最適な照合条件を調べる予定である。今までの実験の感触では、今後この問題に投資するコストにもよるが、文型パターンの適用範囲をあと20%程度向上させる方法はありそうだと思われ。

（４）「文型パターンパーサー」の試作

「文型パターン照合プログラム」の作成経験を元に、「文型パターンパーサー」試作した。このプログラムは、文型パターン検索機能だけでなく、文型パターンを使用した入力文解析機能を実現するものである。現状では、文法属性、意味属性、離散記号、時制・相・様相などに関する様々な指定条件下で、文型パターン辞書から入力された日本文に適合する文型パターンのすべて抽出する機能を実現したが、引き続き、構文情報の取り出し機能を実現する予定である。

（５）意味類型化方式の検討

文型パターンの意味類型化に関する検討としては、因果関係を表す重文に関する意味定義方法の検討から始まって、重文全体の意味分類方法、複文の意味分類方法の検討など例題を使用した机上検討を進め、以下のような基本的な方針を固めた。すなわち、重文複文を構成する節それぞれに対する単文としての意味分類（40種類程度）、従属節の意味分類（100種類程度）、従属節の述部語尾表現分類の３種類の情報を用いて意味分類する方法である。この方法では、すべての文型パターンに対して、これらの分類を表すキーワードは、必須の真理項として付与されるが、それ以外の意味的な特徴を表すキーワードは、任意の真理項として付与される。また、これらの情報付与を半自動的に実行する方法について検討し、一定の見通しを得た。

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，村上仁一，木本泰博：単語意味属性を用いたベクトル空間法,言語処理学会論文誌，Vol.10，No.2，pp.111-128（2003.4）

○ 池原悟，村上仁一，桐沢洋：意味的用法に着目した日本語名詞の英訳語選択について，情報処理学会論文誌，Vol.44，No.5，pp.1343-1353（2003.5）

○ 池原悟，村上仁一，桐沢洋：意味的用法に着目した日本語名詞の英訳語選択について，情報処理学会論文誌，Vol.44，No.5，pp.1343-1353（2003.5）

○ 徳久雅人，守谷有司，池原悟，村上仁一：意味属性の共起による「ＡのＢ」型名詞句の翻訳規則，FIT 札幌，pp.87-89（2003）

○ Yshihiko Nitta, Satoru Ikehara, Naoshi Ikeda, Masahiro Miyazaki, Satoshi Shirai, Katsuyuki Shibata; Reevaluation of the Classic Machine Translation through the Pattern Translation, Proceedings of the PACLING-03, pp.1-4(2003))

○ Satoru Ikehara: A Huge Lexical Database and a new Method for Machine Translation, EAJS Conference, Warsow Poland (2003.8.27-28)　（Invited Lecture)

（２）特許出願

なし