自然言語処理 Vol.3 No.1 編集後記

このコーナーでは，本号に掲載された論文の内容をてっとり早く知りたい方のために，論文の内容，その論文の背景にある学問，技術の動向などをまとめる．

「英語名詞の多義性解消における文脈としての場面情報ぼ評価」(角田，田中) は，英文における名詞である単語の意味を同定する研究である．単語の意味は，例えば bucks という名詞では牡鹿や金銭を含み 7 通りもある．実際の文で bucks がどの意味になるのかという単語の意味同定は自然言語理解においては古くから研究されてきたテーマであるが，依然として困難な問題と認識されている．解決策としては動詞の要求する目的語などの名詞句の意味の制限などを利用する方法が従来採られていた．しかし，この論文ではむしろ目を外へ向け，その文がどのような場面で使用されたかにより単語の意味を同定しようとしている．例えば，賭事の場面を記述する文なら bucks は金銭，狩りの場面なら牡鹿と同定しようというものである．論文ではこれによって正しい単語の意味同定までにかかる探索回数が半減したと報告している．

「名詞間の意味的共起情報を用いた複合名詞の解析」(小林，徳永，田中) は，4 文字以上の漢字からなる複合名詞の構造を解析する手法にについて述べている．ここで言う構造とは複合名詞を構成する複数の基本単語への分解，および基本単語の間の係り受け関係である．このような構造など認識しなくても複合名詞を全部辞書に登録してしまえばよいという意見もあるであろうが，漢字複合語の場合 1,2 語からなる基本語に比べてそれらを組み合わせた複合語の方が圧倒的に多いこと，また複合語は日々新しいものが現れ全てを網羅することが原理的に困難であること，などにより基本語がどうのような構造で組み合わされたかを機械的に認識することは重要である．(例えば，「日米自動車交渉」などは最近の造語ではないだろうか) この研究では 16 万語の 4 文字複合語から基本語ないしはその意味の組合せに関する頻度分布を求め，新しい複合語にはこの頻度分布への適合度の高い分解方法および係り受け構造を求める．さらに，接頭語，数詞などがつく場合もヒューリスティックを導入して処理している．解析精度は 93% と報告されている．

ところで，これら二つの論文で使用されている方法論を吟味すると，処理の対象の単語に一致する単語が必ずしも基本的データの中にない場合がある．例えば，角田らの論文では，場面とそこに現れる単語の連想ペアを予め準備しておく必要がある．しかし，処理対象の文の単語が連想ペアに登録されていないことは頻繁に起こるであろう．また，小林らの論文でも，元々の 16 万語から抽出した個々の頻度分布の中に処理したい複合語に含まれる基本単語がぴったり一致しないことがある．これは，コーパスなどの大量データを使うにも係わらず本当に欲しいデータが案外少ないという自然言語処理研究者を悩ませる問題のひとつである．この問題に対してよく使われる解決策は，狙った単語 (あるいはより一般的に言語現象) そのものがなくても，それに類似の単語 (言語現象) で置き換えるという方法である．では，類似の単語などをどうやって知るか? このとき役に立つのが単語の概念の上下関係を記述するシソーラスである．例

えば，角田の連想ペアの一方を単語そのものではなくシソーラス上の上位概念で置き換えておけばよい．すなわち，狩りの場面に「動物」という概念をペアにすれば，文に牡鹿が現れようと，狐が現れようとそこそこに対処できる．ただし，単語そのものではなく一段抽象化した概念だから，処理の精度が若干落ちるが，これをどのように捌くかが研究者の腕の見せ所である．このように役に立つシソーラスでありながら，国内においてはなかなかよいものがない．言語情報処理の基盤になるインフラストラクチャーであるから，是非とも国や大きな機関が人と金を投入して整備してもらいたい．

「Towards a Linguistic Treatment of Compounds in a Machine Tarnslation Environment」(Sophia Ananiadou) も複合語に関する研究論文である．ただし，Eutrora プロジェクトの一環の研究らしく，対象は英語，ギリシア語，イタリア語などにわたる．また，その内容は，複合語 (Compounds) が形態素的概念なのか，構文的概念なのかを議論するような言語学ベースのものである．どちらがよいかについては過去にいろいろな議論があり論文中でサーベイされている．形態的に 1 単語と見なすかどうかの基準として述べられているが，例えば形容詞を中間に挿入できるかというテストなどは面白い．例えば，surface energy を surface free energy とは言える apple pie を apple big pie とは言えない，というようなテストである．後半では，著者の立場すなわち複合語の構成は基本的には構文的なものであること，しかるに処理システムを考えると，形態素解析と構文解析の中間に複合語解析モジュールを挿入する方法が望ましいことが述べられている．前記の小林の論文と似たテーマを扱っており，複合語の処理の問題は，今後の実用的な自然言語処理に向けてのひとつの基本問題であることがうかがわれる．

「名詞の指示性を利用した日本語文章における名詞の指示対象の推定」(村田，長尾) は，自然言語処理における基本的問題のひとつである名詞の指示対象同定に対する工学的アプローチである．この問題の解決が言語理解や機械翻訳の基礎になることは明らかであり，国内外において自然言語処理の分野で多くの研究成果が発表されている．いわゆる centering 理論もそのひとつである．本誌「自然言語処理」においても因果性を利用する西澤・中川の研究 (2 巻 1 号)，マニュアルという特定の対象において，その特性を利用する森・中川らの研究 (2 巻 3 号) などが掲載されている．村田らの研究は，これらと異なり名詞の指示性という非常に一般的な概念に基づく方法である．英語では a と the の差で指示性が陽に表現されるが，村田らは，助詞，連体詞，所有者表現，時制，同一修飾語の使用などを幅広く利用し日本語の名詞の指示性を求めている．それらのデータを総合的に使用して名詞の指示対象をある種のスコアによって定量的に推定しようというものである．限られたテキストとはいえ，テストサンプルにおいて提案した方法は 80%程度を正しく推定している．このような推定は従来，一般的方法だと 6,7 割の精度であったことに比べると，村田らの結果は注目に値する成果ではないだろうか．

中川裕志 (横浜国立大学)