意味情報処理と概念記述言語CDL

特許分野への複合語処理の応用

慶應義塾大学環境情報学部教授
石崎　俊

PROFILE 1970年東京大学工学部計数工学科卒、工博。同助手を経て、1972年通産省工技院電総研勤務。1985年推論システム研究室長、自然言語研究室長を経て、1992年から慶應義塾大学環境情報学部教授、現在に至る。日本認知科学会会長、言語処理学会会長を歴任。現在、情報処理学会情報規格調査会会長、環太平洋計算言語学国際会議会長。2007年度から特許版産業日本語委員会／技術用日本語プラットフォーム委員会委員

慶應義塾大学SFC研究所上席所員
内山　清子

PROFILE 1997年慶應義塾大学大学院政策・メディア研究科修士課程終了、1998年学術情報センターCOE研究員、2003年スタンフォード大学CSLI訪問研究員、2005年学術博士（慶應義塾大学）、2005年から2008年3月慶應義塾大学大学院政策・メディア研究科特別研究教員准教授
特許版産業日本語委員会委員

1 はじめに

特許分野において、意味に重点をおいた自然言語処理技術を応用することについて考察し、概念記述言語CDLや複合語の処理について述べる。

2 意味情報処理

言語情報について関係を記述する対象の単位としては語、句、節、文などがあり、それらの相互の関係を記述するとき、文法的な関係を記述する場合と、意味的な記述をする場合がある。文法的な関係では、形容詞が名詞を修飾するように、修飾関係または係り受け関係などがある。また、たとえば英語の主語を指す主格という言い方もあり、日本語では助詞の名称をそのまま使用する「が格」、「を格」などの言い方もあって、それらを表層格という。

一方、意味的な関係を表すには、深層格を用いるやり方がある。そもそもはフィルモアの深層格が有名であるが、わが国でも1980年代後半からEDRで研究された深層格があり、フィルモアの深層格よりも実用的にかなり数の多いセットが提案されており、最近では概念記述言語CDLとして受け継がれている。

3 概念記述言語CDL

このCDLは人間が理解するレベルに近い内容で意味記述を定義しているので、表層的な関係ではなく、実際の関係を正確な意味表現で表すものということができる。しかし、数10個の深層格ですべての意味関係を記述することになるので、一つの深層格が担う多様な意味関係を統一的に扱うのは場合によっては難しくなる。色々な文例に深層格を適用して当てはめると、人によって当てはめる深層格が揺らぐ場合が若干ある。

そこで、オーサリングツールなどを使用して、表層格と深層格の関係付けのヒントを与えて、人間が効率的に深層格を割り当てることができるように、深層格を扱うときの制約を使用することが望ましい。

このような深層格を複合語における語基間の関係記述に使用することによって、複合語の内部意味構造を正確に記述することが期待される。さらに、複合語間の関係にも使用できる。

また、今後の研究ではあるが、意味を定量化して扱うことができれば、従来とは異なる方法で最適な深層格の割り当てができる可能性がある。著者の研究室では基本語に関する大規模な連想実験を続けており、深層格に基づく連想語を整理・電子化して連想概念辞書としてまとめ、一般に無料で公開している。

CDLはW3Cを通じて国際標準化を進めており、この

ような意味記述が国内だけでなく国際的に通用する記述形式として広く使用される可能性を持っている。

また、別の国際標準化組織にISO/TC37がある。これは、国際標準化機構ISOの下にあるTC37であり、言語資源の記述形式の標準化を進めている。今後はオントロジーも含めてこの組織での標準化の対象になる可能性がある。

4 複合語の意味構造の分析と特許文の解析

効率的な特許文書検索にとって重要な役割を果たすのは複合語ということが出来る。新しい技術情報を扱う特許文書では特に新しい複合語が現れやすく、検索でキーワードになるのは対象分野の専門用語がほとんどであり、特許文書では同じ意味内容でも異なる表現が用いられる場合が多い。複合語を単語列からなる句の形式で言い換えることや、類義語や関連語を追加するなどの方法が有効である。

そこで、Japioを中心に研究を進めてきたプロジェクトの研究の一環を紹介する。具体的には、（1）複合語を構成する語基間の関係解析、（2）複合語間の関係抽出の試みについて述べる。

（1）語基間の関係解析

特許文書に含まれる複合語は、学術論文や技術雑誌に掲載される専門用語とは異なる場合がある。たとえば、「形態素解析システム」を「形態素を解析するためのシステム」と表現することがある。比較的新しく発展した分野の事項で、専門用語として一般的に普及していない場合にこのような句や文の形式で表現することがある。その場合、複合語を検索キーワードにしても検索漏れが生じてしまう。そこで、複合語を構成する語基間の意味関係を明記しておけば、言い換えによって検索キーワードを拡張することが可能になる。

複合語の語基間の意味関係を決定するためには、語基の形態素、文法と意味に関するそれぞれの属性が重要な手がかりとなる。まず形態素属性として名詞を対象とするとき、漢語と和語に分けて考慮すれば、漢語では従来の品詞体系の枠組みにおける名詞（サ変名詞）やナ形容詞語幹に分類される。しかし、名詞の中でも文中において主語や目的語として用いられる語と直後の語に係る修飾的な役割を持つ語とでは文法的属性が異なる。この異なる属性が意味関係の決定に重要な役割を果たすと考えられる。たとえば、「情報検索」という複合語では、「情報」は目的語や主語となることができるが、動作的な属性は持たない。一方、「検索」は主語となることが出来ると同時に、「する」を後続して動詞「検索する」として用いられることがある。この場合は、この語基間には主語述語関係が成り立つとして「情報を検索する」と言い換えることが可能である。このように語基の文法属性から語基間の意味関係をある程度推定できる。

ここで、実際に複合語をひとつの語として検索する場合と、語基に分けてそれらをAND検索する場合、語基を用いた言い換え表現によって検索する場合を簡単な実験で、適合率と再現率を用いたF値で比較すると、上記の3種の順番で精度が格段に上がることが確認されている。

（2）複合語間の関係抽出

検索キーワードを拡張するためには、上記のような言い換え表現の他に、複合語間の意味関係を抽出するための手がかりとして、複合語に挟まれている定型表現を利用する手法が考えられる。日本語定型表現を用いた方法として、「A などのB」「Aのような B」「Aを含んだB」「Aを用いたB」と表現されるパターンを抽出することによって上位下位関係などを抽出することがある。特許文書においてもこの方法が利用できる。

まず、対象とする複合語と、それらの複合語に挟まれている定型表現のパターンを見つける。対象とする複合語はtf.idfなどの指標により上位にランクされた語から任意の数の語を取り出す。次に定型表現パターンをいくつか決定し、そのパターンが対象複合語の間に含まれているかどうかをチェックする。その結果、対象とする複合語を関連付けることが可能となる。

以上のような複合語の誤基間の関係や複合語間の関係記述に将来はCDLを使用することを考えている。

5 おわりに

意味処理や情報検索で重要な基盤が辞書であり、それらのシステムの性能は使用する辞書で決まるといっても過言ではない。特に、CDLを用いて意味を扱うには、そのために必要な意味情報を収集して使用可能な辞書の形式で電子化することがポイントであり、今後の研究課題ということが出来る。