河　原　大　輔（松山教授）

「Automatic Construction of Japanese Case Frames for Natural Language Unders tanding」
（自然言語理解のための日本語格フレーム自動構築）
平成17年７月25日授与

文章で表される内容は、図１のように、もともとネットワーク的な複雑な関係性をもっている。しかし、それを１次元文字列の文章として表現する際に、人間にとって明らかなことは伝達の効率性から明示的に表現されない。このような関係を復元することが、計算機による自然言語理解の重要なステップとなる。

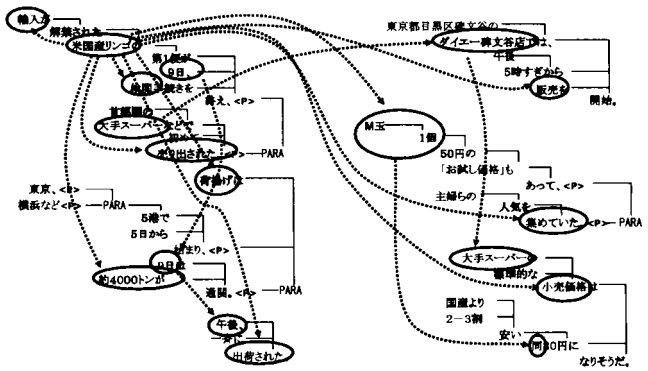

図１．文章中の様々な関係性

文章中の明示されない関係として、例えば図１には「リンゴ→売り出された」という関係がある。このような関係を解析するには、文法程度の知識では不可能であり、「{会社、メーカー、店、…} から {商品、品、野菜、…} が売り出される」のような常識的な知識が必要となる。このような知識は、語がどのような語と関係をもつかという関連性に関するものであり、格フレームと呼ばれる。これまで、重要な用言の典型的な格フレームについては、人手でつくるということが試みられてきた。しかし、用言の数は膨大であり、そのような関係を人手で書き尽くすには非常に大きなコストがかかる。さらには、日々生まれてくる新語や、ドメインごとの専門用語が存在するため、人手による格フレーム作成には限界がある。

本研究では、大規模テキストを自動的に解析することによって格フレームを構築する手法を提案する。自動解析としては、現在のところ実用的な精度を実現している構文解析を用いることができるが、単純に大規模テキストに構文解析を施して、その解析結果を利用するだけでは、幅広い関係を含むような格フレームを得ることはできない。なぜなら、構文解析から得られる関係は当然構文関係でしかなく、それ以上の関係の情報は文章理解を行わないと得られないからである。つまり、文章理解には格フレームが必要であり、一方、格フレーム獲得には文章理解が必要というデッドロックに陥ってしまう。そこで、本研究では、格フレーム構築と文章理解を漸進的に進めることにより格フレームを構築することを考案した。まず、構文解析を行うことによって、「が」「を」などの格助詞が付属している基本的な用例を収集し、第一段階の格フレームを得る。次に、その格フレームを利用した格解析を行うことによって、係助詞句「～は」や被連体修飾詞に関する関係を得、格フレームを高度化する。

このような漸進的処理によって、新聞記事約2600万文から格フレームを構築した。構築された格フレームには、約18,000個の用言が含まれており、１用言あたりの平均格フレーム数は約17.9個である。この格フレームを人手および構文・格・省略解析を通して評価を行い、その結果、高精度かつ実用的なものが構築されていることを確認した。

なお、本研究は、黒橋禎夫先生（当時、東京大学大学院情報理工学系研究科助教授、現京都大学大学院情報学研究科教授）の指導のもとで行った。お世話になった先生方に感謝の意を表したい。