平成１５年度　情報工学コース卒業研究報告要旨

外山 研究室	氏　　名	萩　原　正　人
卒業研究題目	クロスリンガル情報検索のための多言語シソーラス自動構築に関する研究	

近年、世界のボーダーレス化、インターネットの多言語化が進む中で、ある言語で記述された質問文によって別の言語で記述された情報を検索するクロスリンガル情報検索の必要性が高まっている。クロスリンガル情報検索において多言語シソーラスは重要な役割を果たす。シソーラスとは語を同義語・反義語などの意味情報に従って分類・整理した語彙集のことである。多言語シソーラスは複数の言語の語彙を含むシソーラスであり、クロスリンガル情報検索の性能向上に貢献したり、言語に依存しない意味表現である中間言語として用いたりすることができる。しかし、多言語シソーラスを人手で構築することには、コストの問題や、分類基準の言語間での統一などの問題が常に存在する。

それに対して、多言語シソーラスを自動構築する研究が行われてきた。自動構築のためには、対訳コーパスと呼ばれる、対訳関係にある文書の対を大量に集めた文書データを知識源とする手法が一般的である。しかしこの手法には、対訳コーパスが簡単には得られない、同じ言語内での語の関係に関する情報が得られないなどの問題がある。

そこで本研究では、より容易に入手可能な知識源である辞書を利用することによって、多言語シソーラスを自動構築する手法を提案する。具体的には、国語辞典 (日本語-日本語辞典) と英英辞典の記述を用いて、日英２言語シソーラスを自動構築する。

これまで、辞書から単言語シソーラスを自動構築する手法はさかんに研究されてきた。それらの研究では、各索引語が辞書の語義定義文中に出現する頻度を用いて、シソーラスに含める見出し語をそれぞれベクトル表現する。このようにすると、見出し語間の類似度が、対応するベクトルの内積などの尺度によって定量的に求められる。本研究はその多言語への拡張であると位置づけることができる。

本研究では、日本語と英語の語を同一の特徴量を用いてベクトル表現するという方法をとった。そのための特徴量として、定義語ペアを用いた。定義語ペアとは、例えば (“政府”, “government”) のように、日本語と英語の索引語のうち、ほぼ対等関係にあるものを組にしたものである。この定義語ペアは日本語・英語共通の索引語として扱うことができる。そのため、単言語の場合と同様に、辞書の語義定義文中での出現頻度を用いて、各見出し語をベクトル表現することができる。その後、主成分分析の一種である LSI 法を用いて主要な意味を抽出し、ベクトル空間上に各見出し語を配置する。これによって、日本語・英語の別にかかわらず任意の見出し語の間で類似度が求められるようになる。なお、定義語ペアは、日本語と英語の索引語の集合から、和英辞典と英和辞典の記述を用いて、対訳として適切なペアを選択するという方法で自動生成した。

以上の手法を用いて、Longman Dictionary of Contemporary English 3rd ed. の主要見出し語 (3508 語) と国立国語研究所による教育基本語彙 (2039 語) を対象にして日英２言語シソーラスを自動生成した。その結果、類似する語を求めることができ、本手法と、それによって構築された多言語シソーラスの有用性を確認した。