村　脇　有　吾（黒橋教授）
「Automatic Acquisition of Japanese Unknown Morphemes」
（日本語未知語の自動獲得）
平成 23 年 3 月 23 日授与

日本語のテキストを対象に情報検索や機械翻訳といった応用処理を実現しようとしたとき、まず問題となるのは、世界の多くの言語と異なり、日本語では語を空白によって区切らないことである。そのため、テキストを語に自動分割するという処理（形態素解析）が前処理として広く用いられている。形態素解析には長い研究の歴史があり、現在主流となっている手法は辞書を用いるものである。辞書には、(1) 動詞「書く」が「書か - ない」、「書き - ます」、「書く」と活用するといった文法知識、および (2)「書く」、「話す」などの個々の語が人手により記述される。この方式は新聞記事を対象とした従来の評価実験では高い精度を達成している。

辞書に基づく形態素解析は、テキスト中に出現する語があらかじめ辞書登録されていることを前提としており、辞書にない語（未知語）の解析を誤りやすいという欠点がある。新聞記事向けに人手で整備した辞書を用いると、例えば、「とう痛」、「卵黄嚢」といった専門用語が頻出する論文や、「ググる」、「ようつべ」といった俗語がでてくるウェブテキストに対して、満足な解析結果が得られない。この問題への対処法として、新たな分野のテキストを解析する際、あらかじめ人手で辞書に語彙登録するということが現在でも広く行われている。本論文は、このようなコストのかかる語彙登録を計算機により自動化する手法を提案している。具体的には、人手により文法規則と基本的な語彙は整備ずみという設定のもと、解析対象のテキストに出現する未知語を獲得し、人手の介在なしに直接解析にフィードバックする。

本論文では、テキストからの未知語の自動獲得という課題を、未知語検出、未知語同定、自動獲得した名詞の意味分類という 3 つのサブタスクに整理し、それぞれに対して解法を提示した。未知語検出タスクは、テキスト中に出現する未知語を検出するタスクである。基本的な語彙は人手により登録ずみのため、テキスト中の未知語は一般に低頻度だが、効率的に、しかも高い再現率で発見する手法を提案した。次に、検出された未知語に対して、(1) 形態レベルでの同定と (2) 意味分類という 2 段階の問題への切り分けを行った。形態レベルの同定では、日本語が持つ形態論（文法）的が利用できることに着目し、従来手法が統計的に信頼できないとして無視していたほどの少数の用例から高精度に同定できることを示した。一方、明確な文法的区別に基づかない意味分類では、構文情報を含めたより広い手がかりを利用する分類手法を提案した。

本論文が自動獲得の対象としたのは、語（形態素）という言語の最小単位である。しかし、言語処理を用いた応用処理を実現するうえで、語だけでなく、より長い意味的まとまりである複合語の認識も重要である。今後は、複合語ついても計算機が自動認識できるように研究を発展させたい。

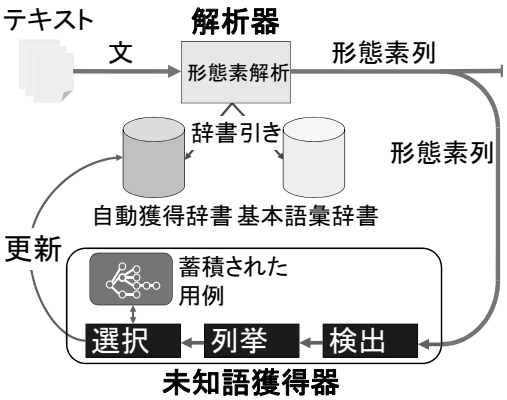

図 1　未知語獲得システム