（ 続紙 １ ）

京都大学	博士（情報学）	氏名	村脇 有吾
論文題目	Automatic Acquisition of Japanese Unknown Morphemes （日本語未知語の自動獲得）		

（論文内容の要旨）

本論文は，日本語形態素解析における未知語問題について論じたものであり，テキストから未知語を自動獲得することにより未知語問題を解決する手法を提案している．また，テキストからの未知語獲得を，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たうえで，それぞれに対して解法を提案しており，全７章から構成されている．

第１章は序論であり，日本語形態素解析における未知語問題という研究背景を紹介したうえで，テキストからの未知語の自動獲得という解決策を提案している．形態素解析用の辞書の構築にこれまでに多くの人的資源を投じられてきたが，この辞書を有効に活用し，人手を新たに介在させることなく未知語を自動獲得するという本研究の方針を示している．未知語の自動獲得は知識獲得問題であり，知識獲得に使うための知識，メタ知識が必要となるが、このメタ知識が，人手で登録された既知語を用いて自動構築できることを論じている．そして，本研究が解くべき３つのサブタスク，未知語検出，未知語同定，自動獲得した名詞の意味分類について，それぞれ具体的にどのようなメタ知識を用いて解くかを概観している．

第２章では，日本語形態素解析について，従来研究を振り返り，人手で整備された辞書を用いる手法が主流となっていることを説明している．また，英語，中国語，トルコ語などの言語との比較を通じて，日本語処理の特性を論じている．そのうえで，辞書を用いた形態素解析においては，辞書にない形態素，未知語が大きな問題となっていることを指摘している．未知語問題に対する従来研究が，形態素解析における未知語処理とテキストからの未知語獲得という大きく二つの流れからなることを振り返り，本研究が解くべき課題をまとめている．

第３章では，テキストから未知語を獲得するという問題設定について論じている．未知語獲得を，テキスト中の未知語の境界を同定し，品詞を割り当てる問題と定義し，境界認定基準と品詞について議論している．後者について，獲得対象を名詞や動詞などのオープンクラスの品詞に限定できることを指摘している．また，日本語の品詞体系が形態論上の区別と意味上の区別の混合であることを指摘したうえで，それに応じて形態論上の獲得と意味上の獲得を区別した段階的な未知語獲得を提案している．さらに，形態論上の獲得をオンラインで行う枠組みを提案し，バッチ処理による従来手法と比較している．

第４章では，未知語検出問題を論じている．まず検出すべき未知語を，既知語によって解釈できない自明な未知語と，既知語の組み合わせによって過分割される未知語の２種類に分類している．そのうえで，後者の過分割された未知語を検出するための手掛かりとして，日本語の表記ゆれの利用を提案している．提案手法は，もし形態素解析結果が正しい分割であれば，その異表記もテキスト中にある程度出現するはずであり，そうでなければその分割結果は未知語の可能性があるという仮定に基づいている．表記ゆれの知識は既知語に対して人手により付与済みであり，既知語と連接する形態素に関する知識はテキストから自動構築できる．提案手法により，ひらがな過分割未知語について，再現率を34.5%から72.0%に向上させることに成功している．

第５章では，検出された未知語に対して，境界を同定し，形態論上の品詞を割り当てる未知語同定問題を論じている．未知語同定問題を候補列挙と候補選択の２段階に分割し，候補列挙の手掛かりとして，日本語の形態論的制約の利用を提案している．テキスト中の個々の未知語用例に対して，形態論的制約を満たす解釈のみを候補とし

て列挙する．これにより，候補選択が，多くの用例を説明できる候補を選択するという単純な手法によって実現でき，さらに未知語獲得がオンライン化できることを示している．形態論的制約の知識はテキスト中の既知語の振る舞いから自動構築でき，人手による整備は必要としない．実験により，少数の用例から97.3-98.4%という高い精度で未知語が獲得され，また獲得の結果，形態素解析の精度が向上することを示している．

第６章では，未知語同定により形態論上の品詞までが付与された状態で獲得された名詞に対して，意味上の区別を行う意味分類問題を論じている．付与すべき意味ラベル間に明確な文法的違いがないことを指摘し，語彙的選好を分類の手がかりとした手法を提案している．分類に有効な語彙的選好は，既知語のテキスト中での振る舞いから自動的に学習される．また，獲得の段階化により，従来研究では用いられなかった構文情報も手がかりとして利用し，高い精度での分類を達成するとともに，構文情報の有効性を実験によって確認している．

第７章は結論であり，本論文を総括している．

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は，日本語形態素解析における未知語問題について，未知語をテキストから自動獲得するという手法と，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未知語自動獲得の３つのサブタスクに関する研究をまとめたもので，得られた主要な成果は以下のとおりである．

１． 未知語検出という従来あまり注目されていなかった問題に取り組み，検出すべき未知語を，既知語によって解釈できない自明な未知語と，既知語の組み合わせによって過分割される未知語の２種類に分類したうえで，後者の過分割された未知語を検出するための手掛かりとして，日本語の表記ゆれの利用を提案した．提案手法により，ベースライン手法では検出が難しかったひらがな過分割未知語について，再現率を大幅に向上させることに成功している．

２． 未知語同定問題に対して，日本語の形態論的制約を手がかりとして用いる手法を提案した．制約に基づく効率的な候補の絞り込みにより，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未知語が獲得できることを示した．また，従来手法がテキストをバッチで処理していたのに対し，提案手法により逐次的に入力されるテキストから未知語獲得するというオンライン処理を実現し，その応用例を示した．

３． 未知語に付与すべき品詞が形態論上の区別と意味上の区別の混合であることを指摘したうえで，形態論上の獲得と意味上の獲得を区別した段階的な未知語獲得を提案した．従来研究がそもそも意味分類を行わなかったり，行っても非常に低い精度に留まっていたのに対し，段階化により問題が単純化することを示した．意味分類においては，従来研究では用いられなかった構文情報も手がかりとして利用した提案手法により，高い分類精度を達成した．

よって，本論文は博士（情報学）の学位論文として価値あるものと認める．また，平成２３年２月２４日実施した論文内容とそれに関連した試問の結果合格と認めた．