論文内容の要旨

博士論文題目　Chinese Synthetic Word Analysis using Large-scale N-grams and An Extendable Lexicon Management System
(大規模なN-gramsデータを用いた中国語合成語解析及び拡張できる語彙管理システム)

氏名　　呂 嘉

(論文内容の要旨)

自然言語には複雑な構造をもつ合成語が多数存在し，特に中国語や日本語のように単語の間にスペース区切りを持たない言語については，合成語の内部構造を知ることが，形態素解析などの基盤的な言語解析にとって重要である．この問題は，機械翻訳などの上位の自然言語処理応用を考える際に，さらに明らかになる．長い合成語が一方の言語の辞書に常に存在するとは限らないため，その内部構造を考慮した翻訳など，柔軟な処理を行うことが重要となる．

本論文では，まず最初に合成語の概念を明らかにし，中国語の単純語と合成語の違いを正確に分類し，定義する．そして，合成語の内部構成要素の統語的関係や形態論的な構造を言語学上の知見に基づいて分類する．次に，中国語の合成語の内部構造を解析するためのいくつかのアプローチを説明する．我々の手法では，合成語内の各部分漢字列が構成要素となりうるかどうかを測るスコアを計算するために，サポートベクターマシンを用いる．さらに，このスコアを用いて内部構造解析を最適の木構造を求める過程として，文脈自由文法の統語解析アルゴリズムであるCKYアルゴリズムを拡張することを提案する．最後に，実験によって中国語合成語の内部構造解析を行い，その結果をいくつかの観点から評価するとともに今後の改良について考察する．

本研究を行うに際して，大規模な中国語合成語の内部構造アノテーションデータを蓄積する必要が生じた．そのため，合成語の内部構造のアノテーションを簡単なマウス操作に行うことができるインタフェースを備え，かつ，利用者が拡張可能な汎用の辞書管理システムの構築を行った．本システム開発の動機について述べ，語彙情報や内部構造情報をどのように表現し利用するかを説明する．最後に，現状のシステム開発の経験を通じて，その欠点や限界を考察し，語彙情報の管理システムをより利用しやすく実現するかについての提言を行う．

氏 名	呂 嘉

（論文審査結果の要旨）

平成22年12月22日に開催した公聴会の結果を参考に平成17年2月18日に本博士論文の審査を行った．以下のとおり，本博士論文は，提案者が独立した研究者として，研究活動を続けていくための十分な素養を備えていることを示すものと認める．

呂 嘉は，本博士論文において，大規模な中国語データを利用して，中国語の合成語の内部構造の解析を行う手法を提案した．また，学習に利用するための合成語データの内部構造のアノテーションを行うことができ，かつ，それを含む様々な語彙情報を格納して管理できる辞書管理システムを開発した．本学位論文の貢献は次のようにまとめることができる．

1. 中国語の単語を，その内部構造と構成要素のつながりを分析し，中国語の合成語の内部構成要素間の統語的，意味的関係の分類を行った．また，構成要素がつながる際に文字の縮約や繰り返しなど，合成語特有に現れる現象についても分類し，合成語解析における問題点を指摘，整理した．また，これにより，合成語解析の研究対象を明確にした．

2. 合成語の内部構造解析のため，構成要素間のつながりを測る尺度について考察し，これを用いて文脈自由文法に基づく統語解析を利用して，合成語の内部構造解析手法を提案した．その際，文字の縮約などの現象を扱うため，従来の上昇型統語解析手法であるCKYアルゴリズムの拡張を行った．また，本提案手法の有効性を実データにより評価した．

3. 自然言語の大規模な語彙データを格納し，様々な情報を付与できる辞書管理システムを開発した．特に，合成語の内部構造アノテーションを簡単なマウスオペレーションで行うことのできるインタフェースを実装し，実験に用いる学習データの構築を効率よく行うことのできる環境を実装した．本システムは，中国語に限らず，言語に依存しない形で実装されており，汎用の辞書管理システムとなっている．

以上のように，中国語の内部構造解析に関して，その解析法だけでなく，実データへのアノテーションやデータの管理を行うシステムの構築を行い，合成語解析に関する汎用性のある手法と環境を構築した本研究は，独創的，かつ，実用的であり，自然言語処理の分野において高い貢献があると評価する．

よって，本論文は，博士（工学）の学位論文として価値あるものと認める．