論文内容の要旨

博士論文題目　Unknown Word Identification for Chinese Morphological Analysis
(中国語形態素解析のための未知語抽出)

氏名　GOH, Chooi-Ling

(論文内容の要旨)

中国語は単語間の境界を明示せずに記述されるため，中国語文の単語分割は計算機による中国語処理において最も基本となる技術である．基本処理としては単語分かち書きだけでなく，個々の単語の品詞同定も重要である．単語の分割に際して，曖昧性を生じる２つの問題がある．一つは分割の曖昧性，もう一つは未知語の出現である．既知語のみにおいても分割の曖昧性は生じるが，未知語の出現については，文脈によりその出現を同定できなければならない．本論文では，未知語処理に重点をおき，機械学習に基づく手法により，それを解決する方法の研究を行った．未知語の出現が同定されたとしても，さらにその品詞を推定する必要がある．本論文では，未知語を構成する文字情報と文脈情報を用いて未知語の品詞を推定する方法を提案した．

未知語同定処理によって得られる未知語候補は，もちろん，すべてが真の単語とは限らない．よって，それらが真に辞書に登録すべき単語がどうかを判定し，そうでない候補を破棄する必要がある．これをすべて人手によって行うには大変な労力を必要とする．本論文では，他に利用可能なタグ付きコーパスを利用することにより，その労力を軽減する方法を提案した．

最初にLDCから公開されているPenn Chinese Treebankを用いて，ベースとなる単語集合(約3万3千語)と統計パラメータを学習し，大規模な生コーパスに提案する未知語処理を適用することにより，多くの未知語候補を抽出した．人手による判定，および，他コーパスでの出現情報等を用いて，最終的に約12万語の中国語辞書を構築した．

中国語の単語認定で問題になるのは，文法あるいは応用の考え方により，単語の分割の仕方が異なる可能性があることである．実際，現在入手可能な中国語タグ付きコーパスの単語分割手法はそれぞれ異なっている．本論文では，２階層の形態素解析モデルを提案した．第１階層では，中国語文を短い単語単位で形態素解析を行う．その後，第２階層において，チャンキング処理を行うことによって，それぞれの分割基準に適合した単語分割に再構成する手法である．

本論文で提案した種々の手法により，中国語の形態素解析を行うための辞書構築，および，解析システム構築を効率よく行うことができることを示した．

氏 名	GOH, Chooi-Ling

（論文審査結果の要旨）

平成18年4月14日に開催した公聴会の結果を参考に平成18年9月1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．

GOH, Chooi-Lingは，本博士論文において，中国語の形態素解析および未知語処理手法を提案し，これらを用いて中国語の辞書を効率よく構築する手法を提案した．

提案された手法とその貢献は次のようにまとめることができる．

1. 中国語文の分かち書き問題において，辞書に基づく前方最長一致と後方最長一致による分割情報を素性として用いる手法を提案し，良好な実験結果を得た．

2. 中国語の形態素解析結果を文字単位に分割し，文字ベースのチャンキングを適用することによって，中国語の未知語を大規模なデータから効率よく検出することができることを示した．

3. 未知語処理を利用し，さらに既存のタグ付きコーパスを参照データとして用いることにより，高い精度で新しい語を辞書に登録する方法を提案した．さらに，実際にその方法を中国語の大規模コーパスに適用することにより，約12万語からなる実用的な中国語形態素辞書を構築した．

4. 中国語形態素解析のための二段階解析法を提案し，分割基準の異なる様々なコーパスにも適用可能な柔軟な手法を示した..

このように，中国語の形態素解析を行うための種々の基礎技術を法を提案し，実システムまで構築した本研究は，独創性と実用性を兼ね備えており，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．