論文内容の要旨

博士論文題目　Biomedical Text Mining Based on Machine Learning: from Information Extraction to Coordination Identification
(機械学習を用いたテキストマイニング —医療情報抽出から並列句解析まで—)

氏名　　原一夫

(論文内容の要旨)

本論文は，臨床試験論文からの情報抽出を目的としている．これは近年一般的になりつつあるコンセプトである「エヴィデンスに基づく医療（EBM)」と密接な関連をもつ．EBMの普及により，医療現場では診断，予後予測，治療，予防に関する最新で正確かつ効果的な方法についての知識が求められるが，それを支援するシステムの作成は人手作業で行われているのが現状である．本研究の最終的な目的は，医療文献を自動的に要約してEBMに必要な情報を患者や医師に提示するシステムの作成であるが，本論文では，その前処理として必要となる，情報抽出タスクと並列句同定タスクについて論じる．

情報抽出タスクでは，既存の自然言語処理の技術を用いてどの程度の精度で重要情報抽出ができるかについて論じる(本博士論文の前半部分)．抽出対象はその臨床試験で比較する治療方法と対象患者である．そこで最初に得る知見は，治療方法と患者を表す基本名詞句の切り出し自体は比較的容易にできることである．しかし同時に，当該臨床試験で比較する治療方法ならびに対象とする患者だけを抽出するのは容易でないことも明らかになる．そこで本論文では文分類によるフィルタリングを試みるが，文の構文構造を素性として用いようとする場合，構文解析の成功が前提となる．しかし，比較結果を記述する臨床試験論文においては，並列句が高頻度に出現する．並列句の存在が構文解析を困難にすることは，自然言語処理学分野ではよく知られている．なおかつ，並列句は情報抽出の観点からも重要な情報を含みやすい．

そこで，本論文の後半部分では並列句の解析手法を新しく提案する．従来手法はルールを発見的に作成するというものがほとんどである．これに対して，我々の提案手法は並列句同定問題を上三角形状の編集グラフにおける系列アラインメントの問題とみなし，編集コスト(素性の重み)を事前に与えることなく，訓練データから学習することができる．GENIAコーパスを用いた実験で，従来手法と比較して良い並列句同定結果を得ることに成功した．なお，提案手法は医薬生物学分野以外のテキストにも適用可能な，自然言語処理の要素技術として用いることができる．

氏 名	原一夫

（論文審査結果の要旨）

平成19年12月27日に開催した公聴会の結果を参考に平成20年2月19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．

原一夫は，本博士論文において，情報抽出技術を利用して臨床試験論文から治療法や試薬の効果や治験結果に関する要約を行うを行う手法を提案し，かつ，それに必要な並列構造解析技術に関する研究を行った．具体的な成果は次の通りである．

1. 臨床試験論文の言語解析を行い，情報抽出技術を適用することにより，臨床試験で比較する治療方法や対象患者に関する情報抽出を行った．基本名詞句の抽出技術と情報抽出のためのパターンを記述することにより，治療法や試薬の効果に関する治験結果をある程度の精度で抽出可能であることを示した．

2. 医学生物学分野の論文では，様々な種類の並列構造が出現する．並列構造の個々の表現は構造的に類似した単語列からなることが多く，系列アラインメントの手法が適用できることが知られている．そのための編集コストを自動的に学習するための新たな手法を提案した．この手法では，完全なアラインメントのアノテーションを行う必要はなく，並列構造の範囲だけを指定するだけでよい．いくつかの実装を提案し，従来手法と比較して精度の高い並列構造解析が可能であることを示した．

臨床試験論文からの治療法や試薬の効果および治験結果に関する情報抽出法を提案し，さらに，統語解析の精度構造に必須の技術である並列構造解析法を提案した本研究は，独創性が高く，しかも実用的であり，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．